

PB-235 302

**EFFICIENT ESTIMATORS FOR SIMULATED
QUEUEING SYSTEMS**

Averill M. Law

California University

Prepared for:

**Environmental Protection Agency
National Science Foundation**

March 1974

DISTRIBUTED BY:

NTIS

**National Technical Information Service
U. S. DEPARTMENT OF COMMERCE
5285 Port Royal Road, Springfield Va. 22151**

BIBLIOGRAPHIC DATA SHEET		1. Report No. ORC 74-7	2. PB 235 302
4. Title and Subtitle EFFICIENT ESTIMATORS FOR SIMULATED QUEUEING SYSTEMS		5. Report Date March 1974	
7. Author(s) Averill M. Law		6. 6. Performing Organization Rept. No. ORC 74-7	
9. Performing Organization Name and Address Operations Research Center 3115 Etcheverry Hall University of California Berkeley, Calif. 94720		10. Project/Task/Work Unit No.	
12. Sponsoring Organization Name and Address Environmental Protection Agency, Washington, D. C. 20460 National Science Foundation, Washington, D. C. 20550		11. Contract/Grant No. R-801870 (EPA) GP-30961X2 (NSF)	
15. Supplementary Notes		13. Type of Report & Period Covered	
16. Abstracts This thesis is primarily concerned with the efficient estimation of mean delay, d , and time average number in queue, Q . For a certain class of queueing systems, it is analytically shown to be more efficient to estimate Q by multiplying an estimate of d by the arrival rate of customers, than to estimate Q directly. This relationship is empirically verified for a much larger class of queueing systems. For single server queueing systems, an estimator is introduced which is more efficient than the normal estimator of d . The efficient estimation of mean waiting time, w , and time average amount of work in system, $E(V)$, is also considered. It is seen to be more efficient to estimate w and $E(V)$ from an estimate of d , than to estimate them directly.		14.	
17. Key Words and Document Analysis. 17a. Descriptors Queueing Simulation Efficient Estimator Regeneration Cycle $L = \lambda w$			
17b. Identifiers/Open-Ended Terms			
Reproduced by NATIONAL TECHNICAL INFORMATION SERVICE U S Department of Commerce Springfield VA 22151			
17c. COSATI Field/Group			
18. Availability Statement Release unlimited.		19. Security Class (This Report) UNCLASSIFIED	21. No. of Pages 97
		20. Security Class (This Page) UNCLASSIFIED	22. Price 4.00

EFFICIENT ESTIMATORS FOR SIMULATED QUEUEING SYSTEMS

by

**Averill M. Law
Operations Research Center
University of California, Berkeley**

MARCH 1974

ORC 74-7

This research has been partially supported by the United States Environmental Protection Agency under Grant R-801870 and the National Science Foundation under Grant GP-30961X2 with the University of California. Reproduction in whole or in part is permitted for any purpose of the United States Government.

DEDICATION

To my wife, Steffi, and daughter, Heather, for the many evenings and weekends they gave up so that I might complete my thesis. Also to my parents and father-in-law for their many years of encouragement.

ACKNOWLEDGEMENT

I would like to sincerely thank my thesis advisor, Professor Ronald W. Wolff, for his many valuable suggestions. I would also like to thank the other members of my committee, Professor C. Roger Glassey and Professor Michael R. Stonebraker, for their interest in my research. In addition, I would like to thank Barbara Brewer for her excellent typing.

ABSTRACT

This thesis is primarily concerned with the efficient estimation of mean delay, d , and time average number in queue, Q . For a certain class of queueing systems, it is analytically shown to be more efficient to estimate Q by multiplying an estimate of d by the arrival rate of customers, than to estimate Q directly. This relationship is empirically verified for a much larger class of queueing systems.

For single server queueing systems, an estimator is introduced which is more efficient than the normal estimator of d .

The efficient estimation of mean waiting time, w , and time average amount of work in system, $E(V)$, is also considered. It is seen to be more efficient to estimate w and $E(V)$ from an estimate of d , than to estimate them directly.

TABLE OF CONTENTS

	PAGE
CHAPTER 1: INTRODUCTION	1
1.1 Estimation of Mean Delay and Time Average Number in Queue	1
1.2 An Alternative Method of Estimation	3
1.3 Asymptotic Properties of Estimators	8
1.4 The Concept of Indirect Estimation.	15
CHAPTER 2: THE M/G/1 QUEUE.	19
2.1 Introduction.	19
2.2 Estimators Using the Total Delay Per Cycle, D_c	20
2.3 Estimators Using the Total Work Per Cycle, V_c	28
2.4 Estimators Using the Total Work in Queue Per Cycle, V_q	33
2.5 Numerical Comparison of the Estimators.	39
2.6 An Additional Estimator of $E(V)$	45
CHAPTER 3: GENERAL SINGLE SERVER QUEUEING SYSTEMS	48
3.1 Introduction and Discussion of the Available Estimators.	48
3.2 Poisson Arrivals in Batches of Size k	49
3.3 Empirical Comparison of $\hat{d}_1$, $\hat{d}_2$, and $\hat{d}_3$	50
3.4 Efficient Estimation of $E(V)$	58
3.5 An Efficient Computer Program for Simulating GI/G/1 Queues	58
CHAPTER 4: MULTI-SERVER QUEUEING SYSTEMS.	64
4.1 Estimation of d and Q	64
4.2 Estimation of w for GI/G/s Queues ($s \geq 1$).	69
CHAPTER 5: COMPARISON OF THE REGENERATIVE AND CLASSICAL METHODS FOR SMALL SAMPLE SIZES	73
5.1 Comparison of $\hat{d}_1$ and $\tilde{d}_1$	73
APPENDIX	
A.1 Moments for the M/G/1 QUEUE	78
A.2 Moments for Single Server Queues with Batch Poisson Arrivals.	83
REFERENCES	87
GLOSSARY	89

CHAPTER 1

INTRODUCTION

1.1 Estimation of Mean Delay and Time Average Number in Queue

The purpose of this thesis is to obtain efficient estimators for steady-state properties of queueing systems, when the data is generated by computer simulation. Two quantities of interest are mean delay and time average number in queue. The delay in queue of a customer is defined to be that period of time from the epoch of his arrival at the system until the epoch at which he commences service. Mean delay in queue, which is denoted by d , is defined as follows:

$$(1.1) \quad d = \lim_{n \rightarrow \infty} \frac{\sum_{i=1}^n D_i}{n} \text{ w.p. } 1,$$

where D_i is the delay in queue of the i^{th} customer and w.p. 1 is the abbreviation for with probability one. The time average number of customers in queue (not including the customer or customers in service), which is denoted by Q , is defined as follows:

$$(1.2) \quad Q = \lim_{t \rightarrow \infty} \frac{\int_0^t N_q(s) ds}{t} \text{ w.p. } 1,$$

where $N_q(t)$ is the number of customers in queue at epoch t .

The conventional way of estimating d is to make a simulation run of length n customers and use the following estimator:

$$(1.3) \quad \tilde{d}_1 = \frac{\sum_{i=1}^n D_i}{n}.$$

Similarly, the conventional way of estimating Q is to make a simulation run of length t time units and use the following estimator:

$$(1.4) \quad \tilde{Q}_1 = \frac{\int_0^t N_q(s) ds}{t}.$$

There are two major disadvantages of the above procedures. First, since d and Q are properties of the steady-state behavior of a queueing system, there is the problem of the initial transient. That is, the initial observations of D_1 or $N_q(t)$ are not representative of steady-state behavior. Of course, this problem could be eliminated if the simulation was started in steady-state. However, if enough information was available to do this, then there would probably be no reason to simulate the system in the first place.

In order to compensate for the bias caused by the initial conditions, many authors have suggested deleting some observations from the estimate. For example, using this procedure d would be estimated from a simulation run of length n customers by

$$(1.5) \quad \tilde{d}_2 = \frac{\sum_{i=m+1}^n D_1}{n - m},$$

where m is the number of observations deleted. Of course, there is the problem of what value of m to use.

Even though this procedure does reduce bias, Blomqvist (1970) has shown that for $M/G/1$ and $GI/M/1$ queues, and large values of n , $m = 0$ is that value of m which minimizes mean square error (MSE), $E\{(\tilde{d}_2 - d)^2\}$.

Another problem with the above method of estimation is determining the accuracy of estimates. The delays of successive customers $D_1, D_2, \dots$ are autocorrelated. If the i^{th} customer has a long delay in queue, it is quite likely that the $i + 1^{\text{st}}$ will also. Thus, the methods of classical statistical analysis for independent identically distributed (i.i.d.) random variables (r.v.'s) are not directly applicable. This has led simulators to resort to such techniques as time series analysis (cf. Fishman and Kiviat (1967)) or the method of batch means (cf. Conway (1963)), both of which are quite subjective.

In this section we discussed the classical method for estimating d and Q . We also pointed out several disadvantages of this method. In the next section we consider an alternative method of estimation which is based on the regenerative structure of most queueing systems. As we shall see, this method of estimation eliminates the problems associated with the classical method of estimation.

In Section 3 we discuss asymptotic properties of estimators. This material provides us with a criterion for comparing two or more estimators of the same quantity.

In Section 4 the central idea of this thesis is introduced. Namely, the concept of indirect estimation. Suppose we want to estimate some quantity θ_1 . We can either estimate θ_1 directly, or we can estimate θ_1 by estimating a second quantity, θ_2 say, to which θ_1 can be related. We shall see that in some cases, indirect estimation results in rather significant variance reductions.

1.2 An Alternative Method of Estimation

In this section we consider an alternative method of estimating d and Q as well as other quantities of interest. We discuss this

method in the context of GI/G/s queues, although it is also valid for other queueing systems such as tandem queues with renewal input, queues with batch Poisson arrivals, etc. Let us begin with a discussion of GI/G/1 queues which we now define. There is a single server with first-in first-out (FIFO) order of service. Let A_i be the time between the arrivals of the i^{th} and $i + 1^{\text{st}}$ customers, and S_i the service time of the i^{th} customer. The sequences $\{A_i\}$ and $\{S_i\}$ are each assumed to be i.i.d., and also to be independent of each other. Define the traffic intensity, ρ , by

$$(2.1) \quad \rho = \frac{E(S)}{E(A)}.$$

It will be assumed throughout this thesis that $\rho < 1$ and $E(A) < \infty$.

Suppose the first customer arrives at epoch 0. He finds the system empty and experiences no delay. Let T_1 be the next epoch a customer arrives to find the system empty. It is a consequence of $\rho < 1$ that T_1 is finite w.p. 1.

The customer who arrives at epoch T_1 finds the system in exactly the same state as the customer who arrived at epoch 0. The behavior of the system from T_1 on is a probabilistic replica of the whole system from epoch 0. The system is said to regenerate itself at epoch T_1 , and the length of time required for this to occur is called a regeneration or busy cycle. The finiteness of T_1 implies the existence of an infinite number of busy cycles. Thus, $\{N_q(t), t \geq 0\}$ is a continuous-time regenerative process which regenerates itself at those epochs when a customer arrives to find the system empty. Similarly, $\{D_n, n \geq 1\}$ is a discrete-time regenerative process which regenerates itself when a customer experiences a zero delay. Furthermore, comparable r.v.'s defined over different busy cycles are

1.1.d. Some examples of r.v.'s of interest are length of a busy cycle, number of customers served in a busy cycle, and total delay of all customers in a busy cycle. These are denoted by B_c , N_c , and D_c , respectively. It is a further consequence of $\rho < 1$ (cf. Ross (1970), page 99) that $E(B_c) < \infty$ and $E(N_c) < \infty$.

Consider now the GI/G/s queue for $s > 1$. There are s servers in parallel with a single queue. We make the same assumptions about the service discipline, the $\{A_i\}$'s, and the $\{S_i\}$'s as we did before. The appropriate definition of ρ is now given by

$$(2.2) \quad \rho = \frac{E(S)}{sE(A)},$$

which we again always assume is less than 1.

The GI/G/s queue ($s > 1$) will regenerate itself when a customer arrives to find the system completely empty. However, the length of a busy cycle no longer need be finite. For example, consider the GI/G/2 queue with $A_i = 2$ and $S_i = 3$ for every i . There is only one busy cycle which is infinite in duration. Whitt (1972) has shown that if $P\{A_i > S_i\} > 0$, then B_c will be finite w.p. 1. We will henceforth assume this condition. Thus, $\{N_q(t), t \geq 0\}$ and $\{D_n, n \geq 1\}$ will once again be regenerative processes. Furthermore, $E(B_c) < \infty$ and $E(N_c) < \infty$. Our conclusions about the GI/G/1 queue now carry over to the GI/G/s queue.

Recall that D_c is the total delay of all the customers in a busy cycle, i.e.,

$$(2.3) \quad D_c = D_1 + D_2 + \dots + D_{N_c}.$$

We would like to show that

$$(2.4) \quad D_c = \int_0^{B_c} N_q(t) dt .$$

We assume, without loss of generality, that the busy cycle begins at epoch 0 and ends at epoch B_c . Let the customers who are served in this busy cycle be numbered 1, 2, ..., N_c and let

$$(2.5) \quad Y_1(t) = \begin{cases} 1 & \text{if the } 1^{\text{th}} \text{ customer is in queue at epoch } t \\ 0 & \text{otherwise} \end{cases} .$$

It is easy to see that

$$(2.6) \quad N_q(t) = \sum_{i=1}^{\infty} Y_i(t) .$$

Furthermore,

$$(2.7) \quad \int_0^{B_c} N_q(t) dt = \int_0^{B_c} \sum_{i=1}^{\infty} Y_i(t) dt = \sum_{i=1}^{\infty} \int_0^{B_c} Y_i(t) dt ,$$

where:

$$(2.8) \quad \int_0^{B_c} Y_1(t) dt = \begin{cases} D_1 & \text{if } 1 \leq N_c \\ 0 & \text{if } 1 > N_c \end{cases} .$$

Equation (2.4) now follows.

It can be shown (cf. Smith (1955)), using only what we have assumed up to now, that

$$(2.9) \quad d = \frac{E(D_c)}{E(N_c)}$$

and

$$(2.10) \quad Q = \frac{\int_0^{\infty} N_q(t) dt}{E(B_c)} = \frac{E(D_c)}{E(B_c)}.$$

Furthermore, if $E(S^2) < \infty$, then it follows from Kiefer and Wolfowitz (1956) that $E(D_c) < \infty$. Thus, d and Q are also finite. We will henceforth assume that $E(S^2) < \infty$.

Equations (2.9) and (2.10) suggest another way of estimating d and Q . Instead of running a simulation until n customers have been served, or until t units of time have elapsed, suppose we run the simulation until exactly m busy cycles have been completed. We have the following data:

$$(2.11) \quad D_{c_1}, D_{c_2}, \dots, D_{c_m};$$

$$(2.12) \quad N_{c_1}, N_{c_2}, \dots, N_{c_m};$$

and

$$(2.13) \quad B_{c_1}, B_{c_2}, \dots, B_{c_m}.$$

Furthermore, we know from previous results that $\{D_{c_i}\}$, $\{N_{c_i}\}$, and $\{B_{c_i}\}$ are each i.i.d. sequences of r.v.'s. Let $\bar{D}_c$, $\bar{N}_c$, and $\bar{B}_c$ be the sample means of these three sequences. It follows that d and Q can be estimated by

$$(2.14) \quad \hat{d}_1 = \frac{\bar{D}_c}{\bar{N}_c}$$

and

$$(2.15) \quad \hat{Q}_1 = \frac{\bar{D}_c}{\bar{B}_c}.$$

Clearly, $\hat{d}_1$ and $\hat{Q}_1$ are consistent estimators of d and Q , respectively.

We will call the method for estimating d and Q that we have just outlined the *regenerative method*. The basic idea behind the regenerative method seems to have originated from Cox and Smith (1961, page 136). However, it was not until recently, cf. Crane and Iglehart (1972 a,b) and Fishman (1972), that it was systematically developed.

For lack of a better name, we will call the original approach to estimation the *classical method*. The classical method attempts to estimate d and Q directly, while with the regenerative method the emphasis is switched to estimating properties of a busy cycle such as $E(D_c)$, $E(N_c)$, and $E(B_c)$. Since observations from all busy cycles are i.i.d., data right from the start of the simulation can be used to estimate the required expectations. Thus, the regenerative method eliminates the problem of the initial transient.

One final note. In the remainder of this thesis regenerative estimators will be denoted by a circumflex (^), and classical estimators by a tilde (~).

1.3 Asymptotic Properties of Estimators

In this section we consider some asymptotic properties of estimators which will be useful when comparing two estimators for the same quantity. This will be made clearer in the next section.

Every estimator based on the regenerative method that we will consider has an analogous estimator based on the classical method,

and vice versa. Furthermore, all estimators when properly normalized are asymptotically normally distributed, and in the case of a regenerative-classical pair of estimators, the variances of their asymptotic distributions (VAD's) are equal. We will show this for $\hat{d}_1$ and $\tilde{d}_1$. The proofs are similar for the other pairs of estimators.

We begin with two lemmas, the first of which is a special case of Theorem 4.4.8 of Chung (1968), and thus will be stated without proof. Let $\xrightarrow{D}$ denote convergence in distribution and let $\xrightarrow{P}$ denote convergence in probability.

Lemma 3.1:

If $X_n \xrightarrow[n \rightarrow \infty]{D} F$, $Y_n \xrightarrow[n \rightarrow \infty]{P} 0$, and $Z_n \xrightarrow[n \rightarrow \infty]{P} 1$, then $(X_n + Y_n)Z_n \xrightarrow[n \rightarrow \infty]{D} F$.

Note that convergence w.p. 1 implies convergence in probability.

Let X_1 be a r.v. defined on the 1th busy cycle of the discrete-time process $\{D_n, n \geq 1\}$ (the X_1 's are comparable r.v.'s). We know from previous results that the sequence $\{X_1\}$ is i.i.d. Furthermore, let $N(n)$ be the number of busy cycles completed by epoch n .

Lemma 3.2:

If $E(X_1) < \infty$, then

$$(3.1) \quad \lim_{n \rightarrow \infty} \frac{X_{N(n)+1}}{n} = 0 \text{ w.p. 1.}$$

Proof:

By the strong law of large numbers (s.l.l.n.),

$$E(X_1) = \lim_{n \rightarrow \infty} \frac{\sum_{i=1}^{N(n)+1} X_i}{N(n)+1} = \lim_{n \rightarrow \infty} \frac{\sum_{i=1}^{N(n)} X_i}{N(n)} - \frac{N(n)}{N(n)+1}$$

(3.2)

$$+ \lim_{n \rightarrow \infty} \frac{X_{N(n)+1}}{n} \frac{n}{N(n)+1} = E(X_1) + E(N_c) \lim_{n \rightarrow \infty} \frac{X_{N(n)+1}}{n} \text{ w.p. 1.}$$

The lemma follows from (3.2).

We are now ready to prove the asymptotic normality of $\hat{d}_1$ and $\tilde{d}_1$. In order to avoid trivialities ($D_1 \equiv 0 \forall_1$), we henceforth assume that $P\{S_n > A_n + A_{n+1} + \dots + A_{n+s-1}\} > 0$, where s is the number of servers. Let $N(0,1)$ denote the distribution function of a normal r.v. with mean 0 and variance 1.

Theorem 3.3:

If $E(D_c^2) < \infty$ and $E(N_c^2) < \infty$, then

$$(3.3) \quad \frac{m^{1/2}(\hat{d}_1 - d)}{\sqrt{V_1}} \xrightarrow[n \rightarrow \infty]{D} N(0,1),$$

and

$$(3.4) \quad \frac{m^{1/2}(\tilde{d}_1 - d)}{\sqrt{V_1}} \xrightarrow[n \rightarrow \infty]{D} N(0,1);$$

where, in (3.4), m is given by

$$(3.5) \quad m = \left\lceil \frac{n}{E(N_c)} \right\rceil \left(\text{greatest integer} \leq \frac{n}{E(N_c)} \right),$$

and

$$(3.6) \quad V_1 = \frac{\text{Var}(D_c - dN_c)}{E^2(N_c)}.$$

Proof:

Note that the conditions of the theorem imply that $V_1 < \infty$. We begin with (3.3). Let $Z_1 = D_{c_1} - dN_{c_1}$. The $\{Z_1\}$'s are an i.i.d. sequence of r.v.'s with zero expectation. Therefore, by the central limit theorem for the partial sums of i.i.d. r.v.'s we have

$$(3.7) \quad \frac{\sum_{i=1}^m Z_1}{m^{1/2} \sqrt{\text{Var}(Z)}} \xrightarrow[m \rightarrow \infty]{D} N(0,1).$$

From the s.l.l.n. we know that

$$(3.8) \quad \lim_{m \rightarrow \infty} \frac{1}{E(N_c)} \frac{\sum_{i=1}^m N_{c_1}}{m} = 1 \text{ w.p. } 1.$$

Therefore, by Lemma 3.1 we get

$$(3.9) \quad \frac{m^{1/2}(\hat{d}_1 - d)}{\sqrt{V_1}} = \frac{mE(N_c)}{\sum_{i=1}^m N_{c_1}} \cdot \frac{\sum_{i=1}^m Z_1}{m^{1/2} \sqrt{\text{Var}(Z)}} \xrightarrow[m \rightarrow \infty]{D} N(0,1).$$

Now consider (3.4). Let

$$(3.10) \quad S_{N(n)} = \sum_{i=1}^{N(n)} N_{c_1}.$$

Then we can write $\sum_{i=1}^n D_i - nd$ as follows:

$$(3.11) \quad \sum_{i=1}^n D_i - nd = \sum_{i=1}^{N(n)} Z_i + \sum_{i=S_{N(n)}+1}^n D_i - d(n - S_{N(n)}) .$$

By the central limit theorem for randomised partial sums of i.i.d. r.v.'s (cf. Chung (1968), page 197) we get

$$(3.12) \quad \frac{\sum_{i=1}^{N(n)} Z_i}{(nE(N_c))^{1/2} \sqrt{V_1}} \xrightarrow[n \rightarrow \infty]{D} N(0,1) .$$

It is easy to see that

$$(3.13) \quad 0 \leq \sum_{i=S_{N(n)}+1}^n D_i \leq D_{c_{N(n)+1}} .$$

Therefore, letting $X_1 = D_{c_1}^2$ in Lemma 3.2, we get

$$(3.14) \quad 0 \leq \lim_{n \rightarrow \infty} \frac{\sum_{i=S_{N(n)}+1}^n D_i}{n^{1/2}} \leq \lim_{n \rightarrow \infty} \frac{D_{c_{N(n)+1}}}{n^{1/2}} = 0 \text{ w.p. } 1 .$$

Similarly,

$$(3.15) \quad 0 \leq d(n - S_{N(n)}) \leq dN_{c_{N(n)+1}} ,$$

and thus

$$(3.16) \quad 0 \leq \lim_{n \rightarrow \infty} \frac{d(n - S_{N(n)})}{n^{1/2}} \leq \lim_{n \rightarrow \infty} \frac{dN_{c_{N(n)+1}}}{n^{1/2}} = 0 \text{ w.p. } 1 .$$

Therefore, by Lemma 3.1 we get

$$(3.17) \quad \frac{n^{1/2}(\bar{d}_1 - d)}{(E(N_c))^{1/2}\sqrt{V_1}} \xrightarrow{D} N(0,1) .$$

The second part of the theorem follows from Lemma 3.1 since

$$(3.18) \quad \lim_{n \rightarrow \infty} \left(\frac{nE(N_c)}{n} \right)^{1/2} = 1 .$$

Note that the argument used in proving (3.4) is similar to one used by Iglehart (1971) in proving a functional central limit theorem for the GI/G/1 queue.

Let θ be some steady-state property of a queueing system that we would like to estimate by simulation. Suppose that θ can be expressed as follows:

$$(3.19) \quad \theta = c \frac{E(X)}{E(Y)} ,$$

where $E(X)$ and $E(Y)$ are properties of a busy cycle such as $E(D_c)$, $E(N_c)$, etc. Suppose we simulate our queueing system for n busy cycles and use the following regenerative estimator to estimate θ :

$$(3.20) \quad \hat{\theta} = c \frac{\bar{X}}{\bar{Y}} .$$

We have the following corollary to Theorem 3.3:

Corollary 3.4:

$\hat{\theta}$ is asymptotically normally distributed and the $VAD(\hat{\theta})$ is given by

$$(3.21) \quad VAD(\hat{\theta}) = \frac{1}{n} \frac{\text{Var}(C\bar{X} - \theta\bar{Y})}{E^2(Y)} = \frac{\theta^2}{n} \left[\frac{E(X^2)}{E^2(X)} - \frac{2E(XY)}{E(X)E(Y)} + \frac{E(Y^2)}{E^2(Y)} \right] .$$

Proof:

The asymptotic normality of $\hat{\theta}$ and the first expression for $VAD(\hat{\theta})$ follow from the proof of (3.3). The second expression for $VAD(\hat{\theta})$ easily follows from (3.19) and the first expression.

In Chapter 2 we will compare two or more estimators for the same quantity on the basis of the VAD's. Since the m in Equation (3.21) will not affect our comparisons, we will henceforth delete it from the expression for the VAD of an estimator.

At the beginning of this chapter we mentioned that there were two disadvantages with the classical method of estimation. Namely, the initial transient and the inability to determine the accuracy of estimates. We later stated that the regenerative method eliminated the problem of the initial transient, but we have yet to consider the problem of how to determine the accuracy of an estimate.

Suppose we use $\hat{d}_1$ to estimate d from a simulation run of length m busy cycles. We now describe a method, which was developed by Crane and Iglehart (1972b), for finding a confidence interval for d . Let $\bar{Z}(m)$ be the sample mean and $S_z^2(m)$ the sample variance of the i.i.d. sequence $\{Z_i\}$. Rewrite (3.7) as follows:

$$(3.22) \quad \frac{m^{1/2} \bar{Z}(m)}{\sqrt{\text{Var}(Z)}} \xrightarrow[m \rightarrow \infty]{D} N(0,1) .$$

However, it follows from the s.l.l.n. that

$$(3.23) \quad \lim_{m \rightarrow \infty} S_z^2(m) = \text{Var}(Z) \text{ w.p. } 1 .$$

Therefore, by Lemma (3.1) we get

$$(3.24) \quad \frac{m^{1/2} \bar{Z}(m)}{S_z(m)} \xrightarrow[m \rightarrow \infty]{D} N(0,1) .$$

Thus, for $0 < \gamma < 1$ and large n we get

$$(3.25) \quad P\left\{-u_{1-\gamma/2} \leq \frac{n^{1/2} \bar{z}(n)}{s_z(n)} \leq u_{1-\gamma/2}\right\} \approx 1 - \gamma,$$

where $u_{1-\gamma/2} = \Phi^{-1}(1 - \gamma/2)$ and $\Phi(\cdot)$ is the distribution function of a $N(0,1)$ r.v. An approximate $100(1 - \gamma)$ percent confidence interval may now be obtained from (3.25) by algebraic manipulation. It should be noted, however, that both the numerator and denominator of (3.25) are functions of d .

Actually the above procedure can also be used to obtain a confidence interval for d when using the classical estimator $\bar{d}_1$. Suppose we run a simulation until n customers have been serviced and use $\bar{d}_1$ to estimate d . It follows from the central limit theorem for randomized partial sums of i.i.d. r.v.'s that

$$(3.26) \quad \frac{(N(n))^{1/2} \bar{z}(N(n))}{s_z(N(n))} \xrightarrow[n \rightarrow \infty]{D} N(0,1),$$

where $N(n)$ is the number of busy cycles completed by each n . A confidence interval for d may now be obtained in a similar manner as before.

1.4 The Concept of Indirect Estimation

Although our main concern in this thesis is the efficient estimation of d and Q , we also consider the estimation of mean waiting in system, w , and the time average amount of work in system, $E(V)$. These latter quantities are defined as follows:

$$(4.1) \quad w = \lim_{n \rightarrow \infty} \frac{\sum_{i=1}^n (D_i + S_i)}{n} \quad \text{w.p. 1}$$

and

$$(4.2) \quad E(V) = \lim_{t \rightarrow \infty} \frac{\int_0^t V(s) ds}{t} \quad \text{w.p. 1.}$$

Suppose we temporarily denote d , Q , w , and $E(V)$ by θ_1 ($i = 1, \dots, 4$), respectively. We shall see that any one of the four quantities, θ_j say, can be expressed as a linear function of any one of the other quantities, θ_k say. That is, there exist constants a and b such that

$$(4.3) \quad \theta_j = a\theta_k + b.$$

Furthermore, let $\hat{\theta}_j$ and $\hat{\theta}_k$ be the direct or natural estimators of θ_j and θ_k , respectively. It follows that we may either estimate θ_j by the direct estimator, $\hat{\theta}_j$, or the indirect estimator $a\hat{\theta}_k + b$. If $\hat{\theta}_j$ is a more efficient estimator of θ_j than $a\hat{\theta}_k + b$, then $\frac{\hat{\theta}_j - b}{a}$ is also a more efficient estimator of θ_k than $\hat{\theta}_k$, and vice versa. Let us consider some examples of this phenomenon. The direct estimator of d is given by

$$(4.4) \quad \hat{d}_1 = \frac{\bar{D}_c}{N_c}.$$

However, it is well known (cf. Little (1961)) that

$$(4.5) \quad Q = \lambda d.$$

Therefore, we may also estimate d by the following indirect estimator:

$$(4.6) \quad \hat{d}_2 = \frac{1}{\lambda} \hat{Q}_1 = \frac{1}{\lambda} \frac{\bar{D}_c}{\bar{B}_c},$$

where $\hat{Q}_1$ is the direct estimator of Q . We shall see that $\hat{d}_1$ is a more efficient estimator of d than $\hat{d}_2$. It follows that $\lambda \hat{d}_1$ is a more efficient estimator of Q than $\lambda \hat{d}_2 = \hat{Q}_1$.

As a second example let us consider the efficient estimation of w . The direct classical estimator of w is given by

$$(4.7) \quad \tilde{w}_1 = \frac{\sum_{i=1}^n (D_i + S_i)}{n} = \tilde{d}_1 + \bar{S},$$

where $\bar{S}$ is the sample mean of the i.i.d. sequence $\{S_i\}$. However, since the moments of the service distribution would have to be known to simulate the system in the first place, we can also estimate w by the following indirect classical estimator:

$$(4.8) \quad \tilde{w}_2 = \tilde{d}_1 + E(S).$$

In effect, we estimate d and add a constant. We will show in Chapter 4, that for all GI/G/s queues ($s \geq 1$), $\tilde{w}_2$ has smaller variance than $\tilde{w}_1$ for all sample sizes n .

If we are going to compare two or more estimators for the same quantity, then we must have some criterion for comparison. There are two cases to consider. For M/G/1 queues, single server queues with batch Poisson arrivals, and M/M/s queues ($s > 1$) we will be able to explicitly evaluate the moments of the r.v.'s D_c , N_c , B_c , etc. In this case our criterion for comparison will be the variance of the asymptotic distribution of an estimator. Recall that for a regenerative-classical pair of estimators, the variances of their

asymptotic distributions are the same. Thus, any conclusion that we may draw about a pair of regenerative estimators that we are comparing, will also be valid for the comparable pair of classical estimators, and vice versa. In the remainder of this thesis we will generally find it convenient to work with regenerative estimators.

For the majority of the queueing systems that we will consider, it will not be possible to analytically evaluate the variance of the asymptotic distribution of an estimator. Our criterion for comparison in this case will be the variance of an estimator, which we will estimate from a large number of independent simulation runs.

We conclude this chapter with a summary of Chapters 2 through 5. In Chapter 2 we consider M/G/1 queues. We analytically show that $\hat{d}_1$ is a more efficient estimator than $\hat{d}_2$, but that a third estimator, which we call $\hat{d}_3$, is more efficient than either $\hat{d}_1$ or $\hat{d}_2$. We also show that it is more efficient to estimate $E(V)$ from an estimate of d , than to estimate $E(V)$ directly.

In Chapter 3 the relationships between $\hat{d}_1$, $\hat{d}_2$, and $\hat{d}_3$ that we found in Chapter 2, are empirically verified for arbitrary GI/G/1 queues and also tandem queues.

In Chapter 4 we analytically show that $\hat{d}_1$ is more efficient than $\hat{d}_2$ for the M/M/2 queue. It is also shown that $\tilde{w}_2$ is more efficient than $\tilde{w}_1$ for all GI/G/s queues.

Chapters 2 through 4 are primarily concerned with asymptotic properties of estimators. The question of whether regenerative estimators or classical estimators are more efficient is not considered. In Chapter 5 $\hat{d}_1$ and $\tilde{d}_1$ are compared for small samples. For the particular queueing system considered, the regenerative estimator $\hat{d}_1$ is seen to be superior.

CHAPTER 2

THE M/G/1 QUEUE

2.1 Introduction

In this chapter we consider the efficient estimation of d , Q , and $E(V)$ for M/G/1 queues. Many of the relationships that we will prove for M/G/1 queues, will be empirically verified for arbitrary GI/G/s queues and also tandem queues in Chapters 3 and 4.

We begin in Section 2 with a discussion of four estimators of d which are based on the total delay per cycle, D_c . It is shown that $\lambda \hat{d}_1$ is a more efficient estimator of Q , than $\lambda \hat{d}_2 = \hat{Q}_1$ (the direct estimator).

In Sections 3 and 4 we consider estimators of d which are based on the total work per cycle, V_c , and the total work in queue per cycle, V_c^q , respectively. It is shown that the estimators based on work are more efficient than the analogous estimators based on D_c . However, there is, from the point of general applicability, a more important reason for studying these estimators. The estimators of d considered in Sections 3 and 4 are actually estimators of $E(V)$ (made possible by the fact that $d = E(V)$ for M/G/1 queues), most of which are valid for more general queueing systems. Furthermore, the estimation of $E(V)$ is of interest in its own right. Thus, these two sections are, in effect, a study of the efficient estimation of $E(V)$.

In Section 5 the estimators mentioned above are numerically compared for several M/G/1 queues. In some cases rather significant variance reductions result from the indirect estimation of d , Q , and $E(V)$.

In Section 6 we use the general relationship $E(V) = \rho d + \frac{\lambda E(S^2)}{2}$ to show that it is more efficient to estimate $E(V)$ from an estimate

of $\hat{d}$, then to estimate $E(V)$ from any of the estimators considered in Sections 3 and 4.

The estimators that we consider in this chapter are compared on the basis of the VAD's. We assume that $E(S^4) < \infty$. This will guarantee the finiteness of the VAD's.

2.2 Estimators Using the Total Delay Per Cycle, D_c

In this section we consider estimators of d which use D_c in the numerator. The direct estimator of d is given by

$$(2.1) \quad \hat{d}_1 = \frac{\bar{D}_c}{\bar{N}_c}.$$

However, since $Q = \lambda d$ we can also estimate d by multiplying an estimate of Q by $\frac{1}{\lambda}$. That is, we can estimate d by the following estimator:

$$(2.2) \quad \hat{d}_2 = \frac{1}{\lambda} \hat{Q}_1 = \frac{1}{\lambda} \frac{\bar{D}_c}{\bar{B}_c}.$$

There are two additional estimators that we would like to consider. Both depend on simple relationships between $E(B_c)$ and $E(B)$. It is well known that for single server queues

$$(2.3) \quad E(B_c) = \frac{E(B)}{\rho}.$$

This suggests estimating d by the following estimator:

$$(2.4) \quad \hat{d}_3 = \frac{1}{\lambda} \left(\frac{\bar{D}_c}{\bar{B}} \right) = E(S) \frac{\bar{D}_c}{\bar{B}}.$$

In addition to (2.3) we also have the following relationship:

$$(2.5) \quad B_c = B + I.$$

However, for M/G/1 queues an idle period has the same distribution as an interarrival time. Thus,

$$(2.6) \quad E(B_c) = E(B) + \frac{1}{\lambda}$$

and we can also estimate d by

$$(2.7) \quad \hat{d}_4 = \frac{1}{\lambda} \frac{\bar{B}_c}{\bar{B} + \frac{1}{\lambda}}.$$

We have four estimators which we would like to compare on the basis of the VAD's. It follows from Equation (3.21) of Chapter 1 that we must compute the first and second moments of the r.v.'s D_c , N_c , B_c , B , $B + \frac{1}{\lambda}$, and also the product moments such as $E(D_c N_c)$, $E(D_c B_c)$, etc. These moments can be computed by conditioning on the service time of the customer who begins the busy period, and also the number of customers who arrive during this service time. The calculations necessary to compute the required moments are quite laborious and not given here. Instead, several representative examples of the above technique and a listing of the required moments are given in the appendix. From these moments we get

$$(2.8) \quad \begin{aligned} \text{VAD}(\hat{d}_1) = & \frac{\lambda^2 E(N_c) E(S^4)}{4} + \frac{\lambda(1+\rho) E(N_c) E(S^3)}{3} \\ & + \frac{5\lambda^3 E^2(N_c) E(S^3) E(S^2)}{6} + \frac{\lambda^4 E^3(N_c) E^3(S^2)}{2} \end{aligned}$$

$$(2.8) \quad + \frac{\lambda^2(1+\rho)E^2(N_c)E^2(S^2)}{4} ,$$

$$(2.9) \quad \begin{aligned} \text{VAD}(\hat{d}_2) = & \frac{\lambda^2 E(N_c) E(S^4)}{4} + \frac{\lambda(1+\rho) E(N_c) E(S^3)}{3} \\ & + \frac{5\lambda^3 E^2(N_c) E(S^3) E(S^2)}{6} + \frac{\lambda^4 E^3(N_c) E^3(S^2)}{2} \\ & + \lambda^2 E^2(N_c) E^2(S^2) , \end{aligned}$$

$$(2.10) \quad \begin{aligned} \text{VAD}(\hat{d}_3) = & \frac{\lambda^2 E(N_c) E(S^4)}{4} + \frac{\lambda(1+\rho) E(N_c) E(S^3)}{3} \\ & + \lambda^3 \left[\frac{4}{3} - \frac{1}{2\rho} \right] E^2(N_c) E(S^3) E(S^2) \\ & + \lambda^4 \left[\frac{5}{4} - \frac{1}{\rho} + \frac{1}{4\rho^2} \right] E^3(N_c) E^3(S^2) \\ & + \frac{\lambda^2(1+\rho) E^2(N_c) E^2(S^2)}{2} , \end{aligned}$$

and

$$(2.11) \quad \begin{aligned} \text{VAD}(\hat{d}_4) = & \frac{\lambda^2 E(N_c) E(S^4)}{4} + \frac{\lambda(1+\rho) E(N_c) E(S^3)}{3} \\ & + \frac{5\lambda^3 E^2(N_c) E(S^3) E(S^2)}{6} + \frac{\lambda^4 E^3(N_c) E^3(S^2)}{2} \\ & + \left[\lambda^2 E^2(N_c) - \frac{\lambda^2}{4} \right] E^2(S^2) . \end{aligned}$$

The VAD's of the four estimators are summarized in Table 2.1 for

TABLE 2.1. VAD'S OF THE ESTIMATORS

ESTIMATOR	COEFFICIENTS OF THE FOLLOWING MOMENTS:				
	$E(S^4)$	$E(S^3)$	$E(S^3)E(S^2)$	$E^3(S^2)$	$E^2(S^2)$
$\frac{\bar{D}_c}{E(S)} \frac{c}{\bar{B}}$	$\frac{\lambda^2 E(N_c)}{4}$	$\frac{\lambda(1+\rho)E(N_c)}{3}$	$\lambda^3 \left[\frac{4}{3} - \frac{1}{2\rho} \right] E^2(N_c)$	$\lambda^4 \left[\frac{5}{4} - \frac{1}{\rho} + \frac{1}{4\rho^2} \right] E^3(N_c)$	$\frac{\lambda^2(1+\rho)E^2(N_c)}{2}$
$\frac{\bar{D}_c}{\bar{N}_c}$	"	"	$\frac{5\lambda^3 E^2(N_c)}{6}$	$\frac{\lambda^4 E^3(N_c)}{2}$	$\frac{\lambda^2(1+\rho)E^2(N_c)}{4}$
$\frac{1}{\lambda} \frac{\bar{D}_c}{\bar{B} + \frac{1}{\lambda}}$	"	"	"	"	$\lambda^2 E^2(N_c) - \frac{\lambda^2}{4}$
$\frac{1}{\lambda} \frac{\bar{D}_c}{\bar{B} \bar{B}_c}$	"	"	"	"	$\lambda^2 E^2(N_c)$
$\frac{\bar{V}_c}{\rho} \frac{c}{\bar{B}}$	$\frac{\lambda^2 E(N_c)}{4}$		$\lambda^3 \left[\frac{4}{3} - \frac{1}{2\rho} \right] E^2(N_c)$	$\lambda^4 \left[\frac{5}{4} - \frac{1}{\rho} + \frac{1}{4\rho^2} \right] E^3(N_c)$	
$\frac{\bar{V}_c}{\lambda} \frac{c}{\bar{N}_c}$	"		$\frac{5\lambda^3 E^2(N_c)}{6}$	$\frac{\lambda^4 E^3(N_c)}{2}$	$-\frac{\lambda^2(1+\rho)E^2(N_c)}{4}$

TABLE 2.1. (continued)

ESTIMATOR	COEFFICIENTS OF THE FOLLOWING MOMENTS:				
	$E(S^4)$	$E(S^3)$	$E(S^3)E(S^2)$	$E^3(S^2)$	$E^2(S^2)$
$\frac{\bar{V}_C}{\bar{B} + \frac{1}{\lambda}}$	"		"	"	$-\frac{\lambda^2}{4}$
$\frac{\bar{V}_C}{\bar{B}_C}$	"		"	"	
$\frac{\bar{V}_C}{\bar{B}} + E(V^S)$	$\frac{\lambda^2 \rho^2 E(N_C)}{4}$		$\lambda^3 \left[\frac{1}{3} + \frac{\rho^2}{2} \right] E^2(N_C)$	$\frac{\lambda^4 (1 + \rho^2) E^3(N_C)}{4}$	
$\frac{\bar{V}_C}{\lambda \bar{N}_C} + E(V^S)$	"		$\lambda^3 \left[\frac{1}{3} + \frac{\rho}{2} \right] E^2(N_C)$	$\frac{\lambda^4 E^3(N_C)}{2}$	$-\frac{\lambda^2 \rho^2 (3 - \rho) E^2(N_C)}{4}$
$\frac{\bar{V}_C}{\bar{B}} + \frac{1}{\lambda} + E(V^S)$	"		"	"	$-\frac{\lambda^2 \rho^2}{4}$
$\frac{\bar{V}_C}{\bar{B}_C} + E(V^S)$	"		"	"	

easy comparison.

We will now prove some theorems which compare the VAD's of the four estimators. The first theorem compares $\hat{d}_2$ and $\hat{d}_4$. A consequence of this theorem is that $\lambda\hat{d}_4$ is a more efficient estimator of Q than $\hat{Q}_1$ (the direct estimator).

Theorem 2.1:

$$\text{VAD}(\hat{d}_4) < \text{VAD}(\hat{d}_2) .$$

Proof:

The first four terms of each expression are identical, and

$$(2.12) \quad \left[\lambda^2 E^2(N_c) - \frac{\lambda^2}{4} \right] E^2(S^2) < \lambda^2 E^2(N_c) E^2(S^2) .$$

The next theorem shows that $\hat{d}_1$ is a more efficient estimator than $\hat{d}_4$. But more importantly, it implies in light of Theorem 2.1, that $\lambda\hat{d}_1$ is a more efficient estimator of Q than $\hat{Q}_1$.

Theorem 2.2:

$$\text{VAD}(\hat{d}_1) < \text{VAD}(\hat{d}_4) .$$

Proof:

We must show that

$$(2.13) \quad \frac{\lambda^2(1 + \rho)E^2(N_c)}{4} < \lambda^2 E^2(N_c) - \frac{\lambda^2}{4} .$$

Rewriting (2.13) we get

$$(2.14) \quad (3 - \rho)E^2(N_c) > 1 ,$$

which is true for all $0 < \rho < 1$ since $E(N_c) = \frac{1}{1 - \rho} > 1$.

The last theorem of this section compares $\hat{d}_1$ and $\hat{d}_3$. It shows that $\hat{d}_3$ is at least as efficient an estimator of d as $\hat{d}_1$, and $\lambda \hat{d}_3$ is at least as efficient an estimator of Q as $\lambda \hat{d}_1$ (which is more efficient than $\hat{Q}_1$). There will be equality of estimators only for constant service. In this case

$$(2.15) \quad \hat{d}_3 = E(S) \frac{\bar{D}_c}{\bar{b}} = E(S) \frac{\bar{D}_c}{b \bar{N}_c} = \frac{\bar{D}_c}{\bar{N}_c} = \hat{d}_1,$$

where b is the constant service time.

Theorem 2.3:

$VAD(\hat{d}_3) \leq VAD(\hat{d}_1)$, with equality holding if and only if $Var(S) = 0$ (constant service).

Proof:

The first two terms of each expression are identical. Therefore, we must show that

$$(2.16) \quad \begin{aligned} & \lambda^3 \left[\frac{4}{3} - \frac{1}{2\rho} \right] E^2(N_c) E(S^3) E(S^2) + \lambda^4 \left[\frac{5}{4} - \frac{1}{\rho} + \frac{1}{4\rho^2} \right] E^3(N_c) E^3(S^2) \\ & + \frac{\lambda^2 (1 + \rho) E^2(N_c) E^2(S^2)}{2} \leq \frac{5\lambda^3 E^2(N_c) E(S^3) E(S^2)}{6} \\ & + \frac{\lambda^4 E^3(N_c) E^3(S^2)}{2} + \frac{\lambda^2 (1 + \rho) E^2(N_c) E^2(S^2)}{4}. \end{aligned}$$

Multiplying (2.16) by $\frac{4}{\lambda^2 E^2(N_c) E^2(S^2)}$ and rewriting we get

$$(2.17) \quad \frac{2\lambda \left[\frac{1}{\rho} - 1 \right] E(S^3)}{E(S^2)} + \lambda^2 \left[-3 + \frac{4}{\rho} - \frac{1}{\rho^2} \right] E(N_c) E(S^2)$$

$$-(1 + \rho) \geq 0 .$$

This in turn can be rewritten as

$$(2.18) \quad \frac{2(1 - \rho)E(S^3)}{E(S^2)E(S)} + \frac{(3\rho - 1)E(S^2)}{E^2(S)} - (1 + \rho) \geq 0 .$$

However, if X and Y are r.v.'s, then we know from Schwarz's inequality that

$$(2.19) \quad E^2(XY) \leq E(X^2)E(Y^2) .$$

Therefore, letting $X = S^{3/2}$ and $Y = S^{1/2}$ we get

$$(2.20) \quad E^2(S^2) \leq E(S^3)E(S) ,$$

which implies that

$$(2.21) \quad \frac{E(S^3)}{E(S^2)E(S)} \geq \frac{E(S^2)}{E^2(S)} .$$

It follows from (2.18) and (2.21) that

$$(2.22) \quad \frac{2(1 - \rho)E(S^3)}{E(S^2)E(S)} + \frac{(3\rho - 1)E(S^2)}{E^2(S)} - (1 + \rho) \geq$$

$$(1 + \rho) \left[\frac{E(S^2)}{E^2(S)} - 1 \right] \geq 0 ,$$

the last inequality holding since $\text{Var}(S) = E(S^2) - E^2(S) \geq 0$.

Furthermore, it is clear from (2.22) that the two estimators are equivalent only if $\text{Var}(S) = 0$.

The theorems that have been proven in this section (and will be proven in the next two sections) are summarized in Figure 2.1.

Let $\hat{\delta}_1$ and $\hat{\delta}_2$ be estimators of d . The following notation is used:

$$\hat{\delta}_1 \longrightarrow \hat{\delta}_2 \text{ means } VAD(\hat{\delta}_1) < VAD(\hat{\delta}_2),$$

$$\hat{\delta}_1 \longrightarrow \hat{\delta}_2 \text{ means } VAD(\hat{\delta}_1) < VAD(\hat{\delta}_2) \text{ for } \rho \geq \sqrt{2} - 1,$$

this relationship was also numerically shown to hold for $\rho < \sqrt{2} - 1$ for several service distributions of interest (see Sections 2.4 and 2.5 for explanation).

An asterisk next to either of these symbols means that the estimators are equivalent for constant service.

2.3 Estimators Using the Total Work Per Cycle, V_c

We now consider estimators of d which are based on the total work per cycle, V_c . We have the following expression for $E(V)$ which is analogous to Equations (2.9) and (2.10) of Chapter 1:

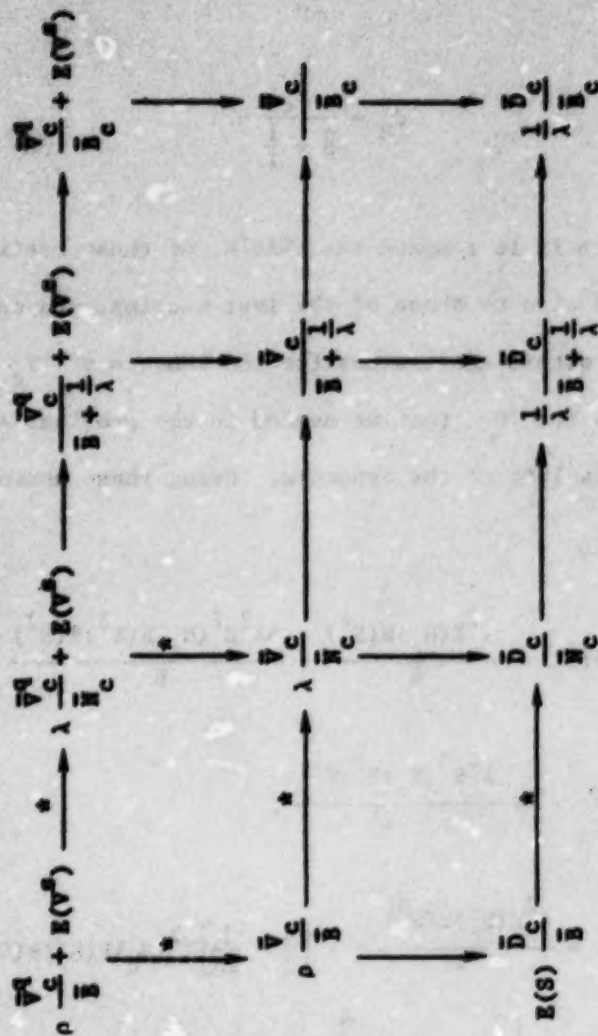
$$(3.1) \quad E(V) = \frac{E \int_0^B V(t) dt}{E(B_c)} = \frac{E(V_c)}{E(B_c)}.$$

Furthermore, it is well known that $d = E(V)$ for M/G/1 FIFO queues.

In light of the above facts, it seems natural to consider

$$(3.2) \quad \hat{d}_5 = \frac{\bar{V}_c}{\bar{B}_c}$$

as an estimator of d . Based on our experience in the previous section, it also seems advantageous to consider the following additional estimators of d :

FIGURE 2.1. RELATIONSHIPS BETWEEN THE ESTIMATORS¹

1 - See page 28 for explanation of symbols.

$$(3.3) \quad \hat{d}_6 = \rho \frac{\bar{V}_c}{\bar{B}},$$

$$(3.4) \quad \hat{d}_7 = \lambda \frac{\bar{V}_c}{\bar{N}_c},$$

and

$$(3.5) \quad \hat{d}_8 = \frac{\bar{V}_c}{\bar{B} + \frac{1}{\lambda}}.$$

Our objective is to compare the VAD's of these estimators to each other and also to those of the last section. In order to do this, we must obtain expressions for the moments of V_c , analogous to ones for D_c that we needed in the previous section. Once again this is left to the appendix. Using these moments, it can be shown that

$$(3.6) \quad \begin{aligned} \text{VAD}(\hat{d}_5) = & \frac{\lambda^2 E(N_c) E(S^4)}{4} + \frac{5\lambda^3 E^2(N_c) E(S^3) E(S^2)}{6} \\ & + \frac{\lambda^4 E^3(N_c) E^3(S^2)}{2}, \end{aligned}$$

$$(3.7) \quad \begin{aligned} \text{VAD}(\hat{d}_6) = & \frac{\lambda^2 E(N_c) E(S^4)}{4} + \lambda^3 \left[\frac{4}{3} - \frac{1}{2\rho} \right] E^2(N_c) E(S^3) E(S^2) \\ & + \lambda^4 \left[\frac{5}{4} - \frac{1}{\rho} + \frac{1}{4\rho^2} \right] E^3(N_c) E^3(S^2), \end{aligned}$$

$$(3.8) \quad \begin{aligned} \text{VAD}(\hat{d}_7) = & \frac{\lambda^2 E(N_c) E(S^4)}{4} + \frac{5\lambda^3 E^2(N_c) E(S^3) E(S^2)}{6} \\ & + \frac{\lambda^4 E^3(N_c) E^3(S^2)}{2} - \frac{\lambda^2 (1 + \rho) E^2(N_c) E^2(S^2)}{4}, \end{aligned}$$

and

$$(3.9) \quad \begin{aligned} \text{VAD}(\hat{d}_8) = & \frac{\lambda^2 E(N_c) E(S^4)}{4} + \frac{5\lambda^3 E^2(N_c) E(S^3) E(S^2)}{6} \\ & + \frac{\lambda^4 E^3(N_c) E^3(S^2)}{2} - \frac{\lambda^2 E^2(S^2)}{4} . \end{aligned}$$

The VAD's of the above four estimators are summarized in Table 2.1.

The following three theorems are completely analogous to those of the last section.

Theorem 3.1:

$$\text{VAD}(\hat{d}_8) < \text{VAD}(\hat{d}_5) .$$

Proof:

The first three terms of each expression are identical. The $\text{VAD}(\hat{d}_8)$ has the additional term $-\frac{\lambda^2 E^2(S^2)}{4}$, which is strictly negative.

Theorem 3.2:

$$\text{VAD}(\hat{d}_7) < \text{VAD}(\hat{d}_8) .$$

Proof:

We must show that

$$(3.10) \quad -\frac{\lambda^2(1+\rho)E^2(N_c)E^2(S^2)}{4} < -\frac{\lambda^2 E^2(S^2)}{4} .$$

Rewriting (3.10) we get

$$(3.11) \quad (1+\rho)E^2(N_c) > 1 ,$$

which is true for all $0 < \rho < 1$.

Theorem 3.3:

$VAD(\hat{d}_6) \leq VAD(\hat{d}_7)$, with equality holding if and only if $Var(S) = 0$ (constant service).

Proof:

It follows from (3.7), (3.8), (2.8), and (2.10) that

$$(3.12) \quad VAD(\hat{d}_7) - VAD(\hat{d}_6) = VAD(\hat{d}_1) - VAD(\hat{d}_3).$$

Therefore, the proof follows from the proof of Theorem 2.3.

We have shown that $\hat{d}_3$ is the most efficient estimator of those considered in Section 2, and that $\hat{d}_6$ is the most efficient estimator of those considered in this section. We now show that $\hat{d}_6$ is a more efficient estimator of d than $\hat{d}_3$.

Theorem 3.4:

$$VAD(\hat{d}_6) < VAD(\hat{d}_3).$$

Proof:

The two expressions have three terms in common. In addition, the expression for the $VAD(\hat{d}_3)$ has two terms which are strictly positive.

For completeness, we state the following three theorems without proof:

Theorem 3.5:

$$VAD(\hat{d}_7) < VAD(\hat{d}_1).$$

Theorem 3.6:

$$VAD(\hat{d}_8) < VAD(\hat{d}_4) .$$

Theorem 3.7:

$$VAD(\hat{d}_5) < VAD(\hat{d}_2) .$$

The theorems that have been proven in this section are summarized in Figure 2.1. We conclude from these theorems that it is always better to estimate d indirectly from an estimate of $E(V)$.

2.4 Estimators Using the Total Work in Queue Per Cycle, V_c^q

In this section we consider estimators of d which are based on the total work in queue per cycle, V_c^q . At any epoch t , the total work in system, $V(t)$, is equal to the sum of the work in queue, $V^q(t)$, and the work in service, $V^s(t)$. Therefore, we have

$$(4.1) \quad E(V) = E(V^q) + E(V^s) ,$$

where $E(V^q)$ and $E(V^s)$ are defined in a similar manner to $E(V)$. However, it is well known (cf. Wolff (1970)) that for GI/G/s queues

$$(4.2) \quad E(V^s) = \frac{\lambda E(S^2)}{2} .$$

Furthermore,

$$(4.3) \quad E(V^q) = \frac{E \int_0^B V^q(t) dt}{E(B_c)} = \frac{E(V_c^q)}{E(B_c)} .$$

Therefore, the following four estimators of d seem worthy of consideration:

$$(4.4) \quad \hat{d}_9 = \frac{\bar{v}_c^q}{\bar{B}_c} + E(v^q) .$$

$$(4.5) \quad \hat{d}_{10} = \rho \frac{\bar{v}_c^q}{\bar{B}} + E(v^q) ,$$

$$(4.6) \quad \hat{d}_{11} = \lambda \frac{\bar{v}_c^q}{\bar{N}_c} + E(v^q) ,$$

and

$$(4.7) \quad \hat{d}_{12} = \frac{\bar{v}_c^q}{\bar{B} + \frac{1}{\lambda}} + E(v^q) .$$

The moments of $\bar{v}_c^q$ necessary to compute the VAD's of the four estimators are given in the appendix. The VAD's are as follows:

$$(4.8) \quad \begin{aligned} \text{VAD}(\hat{d}_9) = & \frac{\lambda^2 \rho^2 E(N_c) E(S^4)}{4} + \lambda^3 \left[\frac{1}{3} + \frac{\rho}{2} \right] E^2(N_c) E(S^3) E(S^2) \\ & + \frac{\lambda^4 E^3(N_c) E^3(S^2)}{2} , \end{aligned}$$

$$(4.9) \quad \begin{aligned} \text{VAD}(\hat{d}_{10}) = & \frac{\lambda^2 \rho^2 E(N_c) E(S^4)}{4} + \lambda^3 \left[\frac{1}{3} + \frac{\rho}{2} \right] E^2(N_c) E(S^3) E(S^2) \\ & + \frac{\lambda^4 (1 + \rho^2) E^3(N_c) E^3(S^2)}{4} , \end{aligned}$$

$$(4.10) \quad \text{VAD}(\hat{d}_{11}) = \frac{\lambda^2 \rho^2 E(N_c) E(S^4)}{4} + \lambda^3 \left[\frac{1}{3} + \frac{\rho}{2} \right] E^2(N_c) E(S^3) E(S^2)$$

$$+ \frac{\lambda^4 E^3(N_c) E^3(S^2)}{2} - \frac{\lambda^2 \rho^2 (3 - \rho) E^2(N_c) E^2(S^2)}{4},$$

and

$$\begin{aligned} \text{VAD}(\hat{d}_{12}) &= \frac{\lambda^2 \rho^2 E(N_c) E(S^4)}{4} + \lambda^3 \left[\frac{1}{3} + \frac{\rho}{2} \right] E^2(N_c) E(S^3) E(S^2) \\ (4.11) \quad &+ \frac{\lambda^4 E^3(N_c) E^3(S^2)}{2} - \frac{\lambda^2 \rho^2 E^2(S^2)}{4}. \end{aligned}$$

The VAD's of the above estimators are also listed in Table 2.1.

We are now in a position to compare the estimators of this section to each other, and also to the estimators of the previous two sections. We would expect the relationships between the four estimators of this section to be similar to what we found before.

This is precisely what the following three theorems show.

Theorem 4.1:

$$\text{VAD}(\hat{d}_{12}) < \text{VAD}(\hat{d}_9).$$

Proof:

Obvious from (4.8) and (4.11).

Theorem 4.2:

$$\text{VAD}(\hat{d}_{11}) < \text{VAD}(\hat{d}_{12}).$$

Proof:

We must show that

$$(4.12) \quad - \frac{\lambda^2 \rho^2 (3 - \rho) E^2(N_c) E^2(S^2)}{4} < - \frac{\lambda^2 \rho^2 E^2(S^2)}{4}.$$

Rewriting (4.12) we get

$$(4.13) \quad (3 - \rho)E^2(N_c) > 1,$$

which is true for all $0 < \rho < 1$.

Theorem 4.3:

$VAD(\hat{d}_{10}) \leq VAD(\hat{d}_{11})$, with equality holding if and only if $Var(S) = 0$ (constant service).

Proof:

The first term of each expression is the same. Therefore, we must show that

$$(4.14) \quad \lambda^3 \left[\frac{1}{3} + \frac{\rho^2}{2} \right] E^2(N_c) E(S^3) E(S^2) + \frac{\lambda^4 (1 + \rho^2) E^3(N_c) E^3(S^2)}{4} \leq \\ \lambda^3 \left[\frac{1}{3} + \frac{\rho}{2} \right] E^2(N_c) E(S^3) E(S^2) + \frac{\lambda^4 E^3(N_c) E^3(S^2)}{2} \\ - \frac{\lambda^2 \rho^2 (3 - \rho) E^2(N_c) E^2(S^2)}{4}.$$

Multiplying (4.14) by $\frac{4}{\lambda^2 E^2(N_c) E^2(S^2)}$ and rewriting we get

$$(4.15) \quad \frac{2\lambda\rho(1 - \rho)E(S^3)}{E(S^2)} + \lambda^2(1 - \rho^2)E(N_c)E(S^2) - \rho^2(3 - \rho) \geq 0.$$

This in turn can be rewritten as

$$(4.16) \quad \frac{2(1 - \rho)E(S^3)}{E(S^2)E(S)} + \frac{(1 + \rho)E(S^2)}{E^2(S)} - (3 - \rho) \geq 0.$$

It follows from (4.16) and (2.21) that

$$\frac{2(1-\rho)E(S^3)}{E(S^2)E(S)} + \frac{(1+\rho)E(S^2)}{E^2(S)} - (3-\rho) \geq \quad (4.17)$$

$$(3-\rho) \left[\frac{E(S^2)}{E^2(S)} - 1 \right] \geq 0.$$

We would now like to prove that $\hat{d}_{10}$ is at least as efficient an estimator of d as $\hat{d}_6$. However, we were only able to prove this for $\rho \geq \sqrt{2} - 1 \approx 0.41$. Fortunately, there is little interest in simulating queueing systems with small values of ρ . In addition, the desired relationship was numerically shown to hold for a number of different services distributions for $\rho < \sqrt{2} - 1$. See the next section for additional comment.

Theorem 4.4:

$VAD(\hat{d}_{10}) \leq VAD(\hat{d}_6)$ for $\rho \geq \sqrt{2} - 1$, with equality holding if and only if $\text{Var}(S) = 0$ (constant service). We exclude from consideration service distributions which have mass at only two points, one of which is 0.

Proof:

We must show that

$$\begin{aligned} & \frac{\lambda^2 \rho^2 E(N_c) E(S^4)}{4} + \lambda^3 \left[\frac{1}{3} + \frac{\rho^2}{2} \right] E^2(N_c) E(S^3) E(S^2) + \frac{\lambda^4 (1 + \rho^2) E^3(N_c) E^3(S^2)}{4} \leq \\ (4.18) \quad & \frac{\lambda^2 E(N_c) E(S^4)}{4} + \lambda^3 \left[\frac{4}{3} - \frac{1}{2\rho} \right] E^2(N_c) E(S^3) E(S^2) + \lambda^4 \left[\frac{5}{4} - \frac{1}{\rho} + \frac{1}{4\rho^2} \right] E^3(N_c) E^3(S^2). \end{aligned}$$

Multiplying (4.18) by $\frac{4}{\lambda^2 E(N_c) E^2(S^2)}$ and rewriting we get

$$(4.19) \quad \frac{(1 - \rho^2)E(S^4)}{E^2(S^2)} + \frac{2(\rho^2 + \rho - 1)E(S^3)}{E(S^2)E(S)} + \frac{(-\rho^2 - 2\rho + 1)E(S^2)}{E^2(S)} \geq 0.$$

For $\rho \geq \sqrt{2} - 1$ the quantity $-\rho^2 - 2\rho + 1$ is nonpositive. It follows from (4.19) and (2.21) that

$$(4.20) \quad \frac{(1 - \rho^2)E(S^4)}{E^2(S^2)} + \frac{2(\rho^2 + \rho - 1)E(S^3)}{E(S^2)E(S)} + \frac{(-\rho^2 - 2\rho + 1)E(S^2)}{E^2(S)} \geq$$

$$(1 - \rho^2) \left[\frac{E(S^4)}{E^2(S^2)} - \frac{E(S^3)}{E(S^2)E(S)} \right] \geq 0.$$

The last inequality follows from a result by Sclove, Simons, and Van Ryzin (1967). Furthermore, equality will hold only if $\text{Var}(S) = 0$.

For completeness we state the following three theorems (which hold for all values of ρ) without proof:

Theorem 4.5:

$\text{VAD}(\hat{d}_{11}) \leq \text{VAD}(\hat{d}_7)$, with equality holding if and only if $\text{Var}(S) = 0$ (constant service).

Theorem 4.6:

$$\text{VAD}(\hat{d}_{12}) < \text{VAD}(\hat{d}_8).$$

Theorem 4.7:

$$\text{VAD}(\hat{d}_9) < \text{VAD}(\hat{d}_5).$$

The theorems that were proven in this section are summarized in Figure 2.1. Except possibly for small values of ρ , we conclude that the estimates of this section are at least as efficient as the analogous estimators of the last section. Furthermore, $\hat{d}_{10}$ is the

most efficient estimator (of d) of the twelve estimators that we considered. Similarly, $\lambda \hat{d}_{10}$ is the most efficient estimator of Q and $\hat{d}_{10}$ is the most efficient estimator of $E(V)$.

2.5 Numerical Comparison of the Estimators

In this section we would like to numerically show how large a variance reduction can be obtained by using an estimator other than $\hat{d}_1$ to estimate d , or an estimator other than $\hat{Q}_1$ to estimate Q . Since we have proven that the estimators using $\bar{B}$ in the denominator are superior, we will restrict our attention to these three estimators, $\hat{d}_1$ (the direct estimator of d), and $\hat{d}_2$ ($\lambda \hat{d}_2$ is the direct estimator of Q).

The relative magnitudes of the VAD's of the above five estimators depend on the variability of the service distribution. Our criterion for the variability of a service distribution is the coefficient of variation, $c.v.(S)$, which is defined as follows:

$$(5.1) \quad c.v.^2(S) = \frac{\text{Var}(S)}{E^2(S)} .$$

The five estimators will be compared for the following four queueing systems which have widely varying $c.v.(S)$'s:

$$(5.2) \quad M/E_4/1, c.v.(S) = .5 ;$$

$$(5.3) \quad M/M/1, c.v.(S) = 1 ;$$

$$(5.4) \quad M/H/1, c.v.(S) = 2 ;$$

$$(5.5) \quad M/H/1, c.v.(S) = 4 .$$

A word of explanation may be necessary for (5.4) and (5.5). The

hyperexponential distribution, which is denoted by H , is defined by the following density function:

$$(5.6) \quad f(x) = p\mu_1 e^{-\mu_1 x} + (1-p)\mu_2 e^{-\mu_2 x} \quad x \geq 0,$$

where $\mu_i > 0$ ($i = 1, 2$) and $0 < p < 1$. In our case we let $p = 0.1$. We can then determine μ_1 and μ_2 from the following relationships:

$$(5.7) \quad \mu_1 = \frac{1 + \frac{9}{c}}{10E(S)},$$

$$(5.8) \quad \mu_2 = c\mu_1,$$

where $c = 7.90$ for (5.4) and $c = 105.77$ for (5.5).

The five estimators were compared for each of the four queueing systems by numerically evaluating the expressions for the VAD's given in Sections 2 through 4. The results of these comparisons are given in Tables 2.2 through 2.5. Consider for example the M/M/1 queue. The variance reductions obtained by estimating d by $\hat{d}_{10}$ rather than $\hat{d}_1$ are 67 and 43 percent for ρ equal to .5 and .7, respectively. Similarly, the variance reductions obtained by estimating Q by $\lambda\hat{d}_{10}$ rather than $\hat{Q}_1$ are 72 and 49 percent for ρ equal to .5 and .7, respectively. Thus, the variance reductions can be quite significant.

In Section 4 we showed that $\hat{d}_{10}$ was at least as efficient as $\hat{d}_6$ for $\rho \geq .41$. In order to see whether this relationship held for $\rho < .41$, we numerically evaluated $VAD(\hat{d}_6)$ and $VAD(\hat{d}_{10})$ for $\rho = .3$. We found that $\hat{d}_{10}$ was more efficient than $\hat{d}_6$ for each of the four queueing systems in this section.

TABLE 2.2. COMPARISON OF THE ESTIMATORS
FOR THE $M/E_4/1$ QUEUE, C.V.(S) = .5.

Ratio \ ρ	.5	.7	.9
$\frac{VAD(\hat{a}_3)}{VAD(\hat{a}_1)}$	.93	.95	.98
$\frac{VAD(\hat{a}_6)}{VAD(\hat{a}_1)}$	.31	.54	.83
$\frac{VAD(\hat{a}_{10})}{VAD(\hat{a}_1)}$	.28	.52	.83
$\frac{VAD(\hat{a}_1)}{VAD(\hat{a}_2)}$	.78	.83	.93
$\frac{VAD(\hat{a}_3)}{VAD(\hat{a}_2)}$	.72	.79	.91
$\frac{VAD(\hat{a}_6)}{VAD(\hat{a}_2)}$	.24	.44	.77
$\frac{VAD(\hat{a}_{10})}{VAD(\hat{a}_2)}$	.22	.43	.77

TABLE 2.3. COMPARISON OF THE ESTIMATORS
FOR THE M/M/1 QUEUE, C.V.(S) = 1.

Ratio \ p	.5	.7	.9
$\frac{VAD(\hat{d}_3)}{VAD(\hat{d}_1)}$	.83	.88	.95
$\frac{VAD(\hat{d}_6)}{VAD(\hat{d}_1)}$	.41	.61	.86
$\frac{VAD(\hat{d}_{10})}{VAD(\hat{d}_1)}$	.33	.57	.85
$\frac{VAD(\hat{d}_1)}{VAD(\hat{d}_2)}$	.85	.89	.96
$\frac{VAD(\hat{d}_3)}{VAD(\hat{d}_2)}$	.71	.78	.91
$\frac{VAD(\hat{d}_6)}{VAD(\hat{d}_2)}$	.35	.55	.82
$\frac{VAD(\hat{d}_{10})}{VAD(\hat{d}_2)}$	.28	.51	.81

TABLE 2.4. COMPARISON OF THE ESTIMATORS
FOR THE M/H/1 QUEUE, C.V.(S) = 2.

Ratio ρ	.5	.7	.9
$\frac{VAD(\hat{d}_3)}{VAD(\hat{d}_1)}$	.76	.82	.93
$\frac{VAD(\hat{d}_6)}{VAD(\hat{d}_1)}$	.61	.73	.89
$\frac{VAD(\hat{d}_{10})}{VAD(\hat{d}_1)}$	.37	.62	.88
$\frac{VAD(\hat{d}_1)}{VAD(\hat{d}_2)}$	.96	.97	.98
$\frac{VAD(\hat{d}_3)}{VAD(\hat{d}_2)}$	.72	.79	.91
$\frac{VAD(\hat{d}_6)}{VAD(\hat{d}_2)}$	.59	.70	.88
$\frac{VAD(\hat{d}_{10})}{VAD(\hat{d}_2)}$	.35	.60	.86

TABLE 2.5. COMPARISON OF THE ESTIMATORS

FOR THE M/M/1 QUEUE, C.V.(S) = 4.

Ratio \ ρ	.5	.7	.9
$\frac{VAD(\hat{d}_3)}{VAD(\hat{d}_1)}$	.65	.76	.92
$\frac{VAD(\hat{d}_6)}{VAD(\hat{d}_1)}$	.58	.73	.90
$\frac{VAD(\hat{d}_{10})}{VAD(\hat{d}_1)}$	.44	.67	.89
$\frac{VAD(\hat{d}_1)}{VAD(\hat{d}_2)}$	.975	.984	.994
$\frac{VAD(\hat{d}_3)}{VAD(\hat{d}_2)}$	.63	.75	.91
$\frac{VAD(\hat{d}_6)}{VAD(\hat{d}_2)}$	.57	.72	.90
$\frac{VAD(\hat{d}_{10})}{VAD(\hat{d}_2)}$	.43	.66	.89

It can be seen from Tables 2.2 through 2.5, that as ρ approaches one, the distinctions between the various estimators diminish. That is, the variance reductions for $\rho = .9$ are smaller than the variance reductions for $\rho = .7$, etc. The reason for this phenomenon can easily be seen from Table 2.1. We have assumed that $E(S^4) < \infty$. Let ρ approach one by holding $E(S)$ fixed, and letting λ approach $\frac{1}{E(S)}$. For large ρ , the dominant term for the VAD of each estimator is the one which contains $E^3(N_c) = \frac{1}{(1-\rho)^3}$. Furthermore, as ρ approaches one, this term will approach the same quantity for each of the twelve estimators.

2.6 An Additional Estimator of $E(V)$

In Sections 3 and 4 we were able to use the concept of work to obtain estimators of d and Q that were more efficient than the analogous estimators using D_c . Our ability to do this depended crucially on the relationship $d = E(V)$. However, there is another way of relating d and $E(V)$ that we would now like to consider. For GI/G/s queues ($s \geq 1$) it can be shown (cf. Brumelle (1971)) that

$$(6.1) \quad E(V) = \rho d + \frac{\lambda E(S^2)}{2}.$$

Recall that the second term on the r.h.s. of (6.1) is just $E(V^S)$. Therefore, (6.1) can be rewritten as follows:

$$(6.2) \quad \rho d = E(V^d).$$

Furthermore,

$$(6.3) \quad E(V^2) = \rho \frac{E(V_c^2)}{E(N)}.$$

This suggests estimating d by the following estimator:

$$(6.4) \quad \hat{d}_{13} = \frac{\bar{V}_c^2}{\bar{N}}.$$

We will now prove that $\hat{d}_{13}$ is never more efficient than $\hat{d}_3$.
Thus, $\hat{d}_6$ and $\hat{d}_{10}$ are both superior to $\hat{d}_{13}$.

Theorem 6.1:

$VAD(\hat{d}_3) \leq VAD(\hat{d}_{13})$, with equality holding if and only if $Var(S) = 0$ (constant service).

Proof:

It follows from (4.9) that

$$(6.5) \quad \begin{aligned} VAD(\hat{d}_{13}) = & \frac{\lambda^2 E(N_c) E(S^4)}{4} + \lambda^3 \left[\frac{1}{2} + \frac{1}{3\rho} \right] E^2(N_c) E(S^3) E(S^2) \\ & + \lambda^4 \left[\frac{1}{4} + \frac{1}{4\rho^2} \right] E^3(N_c) E^3(S^2). \end{aligned}$$

Furthermore, it is easy to show that the statement of the theorem is equivalent to

$$(6.6) \quad \begin{aligned} & \frac{\lambda^2 \rho E^2(N_c) E(S^2)}{E^2(S)} + \frac{\lambda \left[\frac{1}{3} + \frac{5\rho}{6} \right] E(N_c) E(S^3)}{E(S^2) E^2(S)} \\ & - \frac{\lambda(1+\rho) E(N_c) E(S^3)}{3E^2(S^2)} - \frac{\lambda^2(1+\rho) E^2(N_c)}{2} \geq 0. \end{aligned}$$

It follows from the inequality $E(S^2) \geq E^2(S)$ that

$$(6.7) \quad \frac{E(S^3)}{E(S^2)E^2(S)} \geq \frac{E(S^3)}{E^2(S^2)}.$$

Therefore, (6.6) will be true only if the following is true:

$$(6.8) \quad \frac{\lambda^2 \rho E^2(N_c) E(S^2)}{E^2(S)} + \frac{\lambda \rho E(N_c) E(S^3)}{2E^2(S^2)} - \frac{\lambda^2(1 + \rho) E^2(N_c)}{2} \geq 0.$$

This in turn will only be true if the following is true:

$$(6.9) \quad \frac{\lambda^2 E(N_c)}{2} \left[\frac{E(S^3)E(S)}{E^2(S^2)} - 1 \right] \geq 0.$$

The truth of (6.9) follows from (2.20). Furthermore, it is easy to see that the two estimators will be equivalent only if $\text{Var}(S) = E(S^2) - E^2(S) = 0$.

Equation (6.1) did not result in a more efficient estimator of d . On the other hand, it follows from Theorem 6.1 that $\hat{v}_2 = \rho \hat{d}_3 + E(V^S)$ is at least as efficient an estimator of $E(V)$ as $\rho \hat{d}_{13} + E(V^S) = \hat{d}_{10}$, which was the most efficient estimator of $E(V)$ that we considered in Sections 3 and 4. Furthermore, $\hat{v}_2$ is valid for arbitrary GI/G/1 queues. Estimation of $E(V)$ will be further discussed in Section 3.4.

CHAPTER 3

GENERAL SINGLE SERVER QUEUING SYSTEMS

3.1 Introduction and Discussion of the Available Estimators

In this chapter we discuss estimators for single server queuing systems for which the arrival stream is not necessarily Poisson. We begin in this section with a discussion of the available estimators of d . There are essentially three estimators to consider. In Section 2 these estimators are analytically compared for batch Poisson arrivals, and in Section 3 these estimators are empirically compared for arbitrary GI/G/1 queues and also tandem queues. Estimation of $E(V)$ is considered in Section 4. Finally, in Section 5 an efficient computer program for simulating GI/G/1 queues is presented.

In Chapter 2 we essentially considered 12 different estimators of d for M/G/1 queues. Eight of these were actually estimators of $E(V)$ and were made possible by the relationship $d = E(V)$. These estimators are no longer available since this relationship does not hold in general. Furthermore, since an idle period does not have the same distribution as an interarrival time, the estimator $\hat{d}_4$ is no longer available. Therefore, we are left with the following three estimators:

$$(1.1) \quad \hat{d}_1 = \frac{\bar{D}_c}{\bar{N}_c},$$

$$(1.2) \quad \hat{d}_2 = \frac{1}{\lambda} \frac{\bar{D}_c}{\bar{B}_c},$$

and

$$(1.3) \quad \hat{d}_3 = E(S) \frac{\bar{b}}{\bar{b}_c}.$$

3.2 Poisson Arrivals in Batches of Size k

In this section we consider a single server queuing system with arbitrary service distribution, and arrivals of batches, each of size k customers, in accordance with a Poisson process at rate λ . Of course, for $k = 1$ we have the M/G/1 queue. The traffic intensity is now defined as follows:

$$(2.1) \quad \rho = k\lambda E(S).$$

We would like to compare the efficiencies of the estimators $\hat{d}_1$, $\hat{d}_2$, and $\hat{d}_3$, where $\hat{d}_2$ is now defined as follows:

$$(2.2) \quad \hat{d}_2 = \frac{1}{\lambda k} \frac{\bar{b}}{\bar{b}_c}.$$

In particular, we would like to show that the variance reduction obtained by estimating Q by $\lambda \hat{d}_1$ rather than $\hat{Q}_1 = \lambda \hat{d}_2$, can be quite significant for some queuing systems.

We will compare the three estimators on the basis of the VAD's. The moments required to evaluate the VAD's can be computed by conditioning on the service times of the k customers who start a busy period, and also the number of batches of customers who arrive during the time required to service these k customers. The expressions for these moments are given in the appendix.

Since our main objective is to show the magnitude of the variance reduction obtained by using $\hat{d}_1$ rather than $\hat{d}_2$, we will not prove any theorems as we did in Chapter 2. Instead we numerically compared

the VAD's of the three estimators for a single server queueing system with Poisson arrivals in batches of size k and 4-Erlang service. The results of these comparisons are given in Table 3.1 for various values of ρ and k . It can be seen that as the interarrival distribution becomes more irregular (k increases), it becomes more advantageous to use $\hat{d}_1$ rather than $\hat{d}_2$. In some cases variance reductions of more than 40 percent can be obtained. Furthermore, $\hat{d}_3$ is a more efficient estimator of d than $\hat{d}_1$ for all values of ρ and k . However, the variance reduction is a decreasing function of k .

3.3 Empirical Comparison of $\hat{d}_1$, $\hat{d}_2$, and $\hat{d}_3$

In this section we empirically compare $\hat{d}_1$, $\hat{d}_2$, and $\hat{d}_3$ for arbitrary GI/G/1 queues and also tandem queues. Denote the hyper-exponential distribution with coefficient of variation equal to 4 (the same one we considered in Chapter 2) by H_4 . There are 9 possible GI/G/1 queues using E_4 , M , and H_4 as their interarrival and service distributions. For each of these 9 systems we originally made 100 independent runs with $\rho = .5$. For the three systems with exponential interarrival distributions, $E(N_c) = 2$. For these systems we made each run of length 1000 busy cycles or approximately 2000 customers. For the other 6 systems $E(N_c)$ is a function of the interarrival and service distributions. We attempted to make the number of cycles per run such that the total number of customers would be approximately 2000. In Table 3.2 we give the variance reductions (in percent) which resulted from estimating d by $\hat{d}_1$ rather than $\hat{d}_2$. The rows represent the interarrival distributions and the columns the service distributions. It appears that $\hat{d}_1$ is more efficient than $\hat{d}_2$ for each of the 9 systems considered. Furthermore, the greatest

TABLE 3.1. COMPARISON OF THE ESTIMATORS
FOR 4-ERLANG SERVICE.

k	$\frac{VAD(\hat{d}_1)}{VAD(\hat{d}_2)}$			$\frac{VAD(\hat{d}_3)}{VAD(\hat{d}_1)}$		
	$\rho = .5$	$\rho = .7$	$\rho = .9$	$\rho = .5$	$\rho = .7$	$\rho = .9$
1	.78	.83	.93	.93	.95	.98
2	.67	.78	.92	.94	.97	.99
5	.59	.74	.91	.97	.98	.995
10	.56	.73	.91	.98	.99	.997
20	.54	.72	.90	.99	.995	.999

TABLE 3.2. VARIANCE REDUCTION (IN PERCENT) RESULTING FROM
ESTIMATING d BY $\hat{d}_1$ RATHER THAN $\hat{d}_2$, $\rho = 0.5$.

$\begin{matrix} \text{A} \backslash \text{S} \\ \text{A} \end{matrix}$	E_4	M	H_4
E_4	9.7	1.2	0.7
M	24.2	16.2	2.4
H_4	38.4	40.8	21.1

variance reduction occurs for irregular arrivals and regular service. This is consistent with our previous analytical results.

In Table 3.3 we give the variance reductions (in percent) which resulted from estimating d by $\hat{d}_3$ rather than $\hat{d}_1$. As mentioned earlier, we originally made 100 independent runs of each of the 9 systems. Based on these runs, $\hat{d}_3$ was more efficient than $\hat{d}_1$ for 7 out of the 9 systems considered. However, for the $M_4/E_4/1$ and $M_4/M/1$ queues using $\hat{d}_3$ rather than $\hat{d}_1$ resulted in variance increases of 0.4 and 0.9 percent, respectively. This was not consistent with our previous analytical results for batch Poisson arrivals and 4-Erlang service (see Table 3.1). In order to determine whether these variance increases were real or whether they were caused by sampling fluctuations, we made an additional 200 runs of each of the two systems in question.

Using the data from all 300 runs, we now found that $\hat{d}_3$ was more efficient than $\hat{d}_1$ for each of the above two systems. Thus, the results given in Table 3.2 were based on 300 runs for two systems and 100 runs for the other 7 systems. We also used these 200 additional runs to get improved estimates for Tables 3.2 and 3.4 (see Section 3.4). Furthermore, it appears that the greatest variance reduction occurs for regular arrivals and irregular service.

For the queueing systems that we have considered up to now, $\hat{d}_1$ appears to be a more efficient estimator than $\hat{d}_2$ and $\hat{d}_3$ appears to be a more efficient estimator than $\hat{d}_1$. In order to determine the generality of these results, we also simulated several tandem queueing systems. That is, a queueing system where the output of one single server queue is the input to another. A regeneration cycle is now defined to be that period of time between successive epochs when a customer arrives at the first server to find the

**TABLE 3.3. VARIANCE REDUCTION (IN PERCENT) RESULTING FROM
ESTIMATING d BY $\hat{d}_2$ RATHER THAN $\hat{d}_1$, $\rho = 0.5$.**

$\Delta \backslash S$	E_4	M	H_4
E_4	12.1	19.0	39.9
M	7.4	13.9	40.8
H_4	0.8	1.7	26.6

entire system empty.

The first tandem queueing system that we simulated had H_4 for its interarrival distribution and E_4 for each of its service distributions. We made 100 independent runs each of length 100 cycles with $\rho = .5$. We found that the expected number of customers per cycle was 20.5, using $\hat{d}_1$ rather than $\hat{d}_2$ resulted in a 48.1 percent variance reduction, and using $\hat{d}_3$ rather than $\hat{d}_1$ resulted in a 1.5 percent variance increase. There are several reasons to believe that the variance increase was caused by sampling fluctuations. Our remarks will also apply to the $H_4/E_4/1$ and $H_4/M/1$ queues. First, for another system with very irregular arrivals and 4-Erlang service (see Table 3.1 for $k = 20$), we analytically showed that using $\hat{d}_3$ rather than $\hat{d}_1$ resulted in a 1 percent variance reduction. Differences of this magnitude in either direction would be difficult to detect by simulation. Second, it can be seen from Table 3.3 that the variance reduction resulting from using $\hat{d}_3$ rather than $\hat{d}_1$ decreases as the service distribution becomes more regular. In fact, for constant service ($c.v.(S) = 0$) the estimators $\hat{d}_3$ and $\hat{d}_1$ are equivalent. This may be the worst case.

The second tandem queueing system that we simulated had E_4 for its interarrival distribution and H_4 for each of its service distributions. We made 100 independent runs each of length 500 cycles with $\rho = .5$. We found that the expected number of customers per cycle was 4.25, using $\hat{d}_3$ rather than $\hat{d}_1$ resulted in a 36.8 percent variance reduction, and using $\hat{d}_1$ rather than $\hat{d}_2$ resulted in a 1.4 percent variance increase. Once again, there is reason to believe that the variance increase was caused by sampling fluctuations. It can be seen from Table 3.2 that the variance reduction resulting

from using $\hat{d}_1$ rather than $\hat{d}_2$ decreases as the interarrival distribution becomes more regular. In fact, for constant interarrival times the estimators $\hat{d}_1$ and $\hat{d}_2$ are equivalent. This may be the worst case.

Both of the tandem queueing systems that we considered above were extreme cases. We also did a shorter simulation study of the tandem queueing system with all distributions exponential. In particular, we made 100 runs of length 100 cycles with $\rho = .5$. It is known that $E(N_c) = 4$ for this system. We found that using $\hat{d}_1$ rather than $\hat{d}_2$ resulted in a 27.1 percent variance reduction, and using $\hat{d}_3$ rather than $\hat{d}_1$ resulted in a 13.1 percent variance reduction.

Based on our analytical and empirical results, it appears that $\hat{d}_3$ is the most efficient estimator of d , and that $\lambda \hat{d}_3$ is the most efficient estimator of Q .

We conclude this section with some ideas on why $\hat{d}_1$ is a more efficient estimator than $\hat{d}_2$. Suppose a r.v. Z can be written as the product of two other r.v.'s X and Y , that is, $Z = XY$. If $\text{Var}(X)$ and $\text{Var}(Y)$ are sufficiently small, then an approximate expression for $\text{Var}(Z)$ (cf. Lindley (1969), page 134) is given by

$$(3.1) \quad \text{Var}(Z) \approx E^2(Y)\text{Var}(X) + 2E(X)E(Y)\text{Cov}(X,Y) + E^2(X)\text{Var}(Y).$$

We can write $\frac{\bar{D}}{\bar{B}_c}$ as follows:

$$(3.2) \quad \frac{\bar{D}}{\bar{B}_c} = \frac{\bar{D}_c}{\bar{N}_c} \frac{\bar{N}_c}{\bar{B}_c}.$$

If the simulation is run for a large number of cycles, then $\text{Var}\left(\frac{\bar{D}_c}{\bar{N}_c}\right)$

and $\text{Var} \left\{ \frac{\bar{N}_c}{\bar{B}_c} \right\}$ will be small. Therefore, an approximate expression

for $\text{Var} \left\{ \frac{\bar{D}_c}{\bar{B}_c} \right\}$ is given by

$$(3.3) \quad \text{Var} \left\{ \frac{\bar{D}_c}{\bar{B}_c} \right\} \approx E^2 \left\{ \frac{\bar{N}_c}{\bar{B}_c} \right\} \text{Var} \left\{ \frac{\bar{D}_c}{\bar{N}_c} \right\} + 2E \left\{ \frac{\bar{D}_c}{\bar{N}_c} \right\} E \left\{ \frac{\bar{N}_c}{\bar{B}_c} \right\} \text{Cov} \left\{ \frac{\bar{D}_c}{\bar{N}_c}, \frac{\bar{N}_c}{\bar{B}_c} \right\} \\ + E^2 \left\{ \frac{\bar{D}_c}{\bar{N}_c} \right\} \text{Var} \left\{ \frac{\bar{N}_c}{\bar{B}_c} \right\}.$$

Furthermore, $E^2 \left\{ \frac{\bar{N}_c}{\bar{B}_c} \right\}$ will very closely approximate $\left\{ \frac{E(N_c)}{E(B_c)} \right\}^2 = \lambda^2$.

Therefore, the first term on the r.h.s. of (3.3) very closely approximates the variance of $\lambda \hat{d}_1$. In addition, the third term on the r.h.s. of (3.3) is nonnegative. Therefore, if we knew that the approximations were accurate and also that the second term on the r.h.s. of (3.3) was nonnegative, then we would have some idea why $\lambda \hat{d}_1$ has smaller variance than $\hat{Q}_1 = \frac{\bar{D}_c}{\bar{B}_c}$.

Now $\frac{\bar{N}_c}{\bar{B}_c}$ can be considered as an estimate of λ , say $\hat{\lambda}$. If for some simulation run $\hat{\lambda}$ is large (the mean interarrival time is small), then the estimate for d , $\frac{\bar{D}_c}{\bar{N}_c}$, should be large also. Therefore, it is intuitively plausible that the covariance term is positive. For each of the 9 simulated systems, we estimated the r.h.s. and l.h.s. of (3.3). We found the above approximations to be very accurate, and also the second term on the r.h.s. of (3.3) to be positive.

3.4 Efficient Estimation of $E(V)$

In this section we consider the efficient estimation of the time average amount of work in system, $E(V)$, for single server queues. The direct estimator of $E(V)$ is given by

$$(4.1) \quad \hat{v}_1 = \frac{\bar{v}_c}{B_c}.$$

However, because of Equation (6.1) of Chapter 2, we can also estimate $E(V)$ indirectly by the following estimator:

$$(4.2) \quad \hat{v}_2 = \rho \hat{d}_3 + \frac{\lambda E(S^2)}{2},$$

where $\frac{\lambda E(S^2)}{2}$ is a known constant and $\hat{d}_3$ is the estimator of d that we found to be most efficient.

In order to compare $\hat{v}_1$ and $\hat{v}_2$, we originally made 100 independent runs (see Section 3.3 for comment) for each of the 9 queueing systems that we considered in Section 3.3. Actually $E(V)$ was estimated at the same time as d , and thus the run lengths were the same as before. The results of these simulation runs are given in Table 3.4. For each of the 9 systems considered, $\hat{v}_2$ has considerably smaller variance than $\hat{v}_1$.

3.5 An Efficient Computer Program for Simulating GI/G/1 Queues

In this section we present an efficient method (in the sense of the number of computer operations required to simulate a given number of busy cycles) for simulating GI/G/1 queues. Our method is based on the theory of random walks (cf. Feller (1966), Page 189).

Suppose the first customer arrives at epoch 0. Let

**TABLE 3.4. VARIANCE REDUCTION (IN PERCENT) RESULTING FROM
ESTIMATING $E(V)$ BY $\hat{v}_2$ RATHER THAN $\hat{v}_1$, $\rho = 0.5$.**

$\begin{array}{c} S \\ A \end{array}$	E_4	H	H_4
E_4	85.9	84.9	80.4
H	60.9	67.7	84.4
H_4	40.6	46.0	66.1

$$(5.1) \quad X_n = S_n - A_n,$$

where A_n and S_n are defined as in Chapter 1. In addition, let

$$(5.2) \quad Z_0 = 0, Z_n = \sum_{i=1}^n X_i.$$

The first busy period will have $N_c = n$ customers if

$$(5.3) \quad Z_1 \geq 0, Z_2 \geq 0, \dots, Z_{n-1} \geq 0, \text{ but } Z_n < 0.$$

Given that $N_c = n$, we have the following:

$$(5.4) \quad B = \sum_{i=1}^n S_i,$$

$$(5.5) \quad B_c = \sum_{i=1}^n A_i,$$

and

$$(5.6) \quad D_1 = Z_{i-1} \text{ for } i = 1, 2, \dots, n.$$

Furthermore, given that $N_c = n$, the $n + 1^{\text{st}}$ customer will arrive to find the system in exactly the same state as the customer who arrived at epoch 0. The relationships given by (5.3) through (5.6) now apply to the second busy period, etc.

We now present a computer program based on the preceding ideas. Our objective is to estimate d by the following estimator:

$$(5.7) \quad \hat{d}_3 = E(S) \frac{\bar{D}_c}{\bar{B}} = E(S) \frac{\sum_{i=1}^m D_{c1}}{\sum_{i=1}^m B_1},$$

where m is the number of busy cycles that the simulation is to be run. The computer program will use the following variables:

ATM = mean interarrival time,

STM = mean service time,

MAXNBC = the number of busy cycles that the simulation is to be run,

NBC = the number of the busy cycle that is currently being simulated,

SUMD = total delay of all customers, equals $\sum_{i=1}^m D_{c_i}$ at the end of the simulation run,

SUMST = the sum of the service times, equals $\sum_{i=1}^m B_i$ at the end of the simulation run,

RANF = library function that generates $U(0,1)$ r.v.'s,

AT = interarrival time,

ST = service time,

and

D = estimated mean delay.

The computer program, which was written for a CDC 6400, follows:

```
PROGRAM GIG1 (INPUT,OUTPUT)
READ 10, ATM, STM, MAXNBC
10 FORMAT (2F5.0,I5)
C
C *** INITIALIZE SYSTEM
C
NBC = 0
SUMD = 0.
SUMST = 0.
C
C *** BEGIN NEW BUSY CYCLE
C
20 NBC = NBC + 1
SUMX = 0.
C
C *** COMPUTE INTERARRIVAL AND SERVICE TIMES
C
30 R = RANF(0.)
AT = -ATM*ALOG(R)
R = RANF(0.)
ST = -STM*ALOG(R)
SUMST = SUMST + ST
X = ST - AT
SUMX = SUMX + X
C
C *** DETERMINE WHETHER BUSY PERIOD IS OVER
C
IF(SUMX.LT.0) GOT TO 40
```



```

C
C *** ADD DELAY OF THIS CUSTOMER TO TOTAL DELAY
C
      SUMD = SUMD + SUMX
      GO TO 30
C
C *** BUSY PERIOD IS OVER.  DETERMINE WHETHER DESIRED NUMBER
C *** OF CYCLES HAVE BEEN COMPLETED.  IF SO, COMPUTE AND PRINT THE
C *** ESTIMATED MEAN DELAY.
C
      40 IF (NBC.LT.MAXNBC) GO TO 20
      D = (STM*SUMD)/SUMST
      PRINT 50,D
      50 FORMAT(1H1,5X,23HESTIMATED MEAN DELAY = ,F9.3)
      END

```

One final note. The above program was written for an M/M/1 queue. In order to simulate an arbitrary GI/G/1 queue, it is only necessary to change the statements which generate the interarrival and service times.

CHAPTER 4

MULTI-SERVER QUEUING SYSTEMS

4.1 Estimation of d and Q

In this section our objective is the efficient estimation of d and Q for GI/G/s queues ($s > 1$). In Chapter 3 we considered three different estimators of d . The first two of these, which are still valid for multi-server queues, are the following:

$$(1.1) \quad \hat{d}_1 = \frac{\bar{D}_c(m)}{\bar{N}_c(m)}$$

and

$$(1.2) \quad \hat{d}_2 = \frac{1}{\lambda} \frac{\bar{D}_c(m)}{\bar{B}_c(m)},$$

where m is the length of the simulation in busy cycles. The third estimator, which was denoted by $\hat{d}_3$, is not valid for multi-server queues since $E(B) \neq \rho E(B_c)$. Therefore, we will concentrate our attention on $\hat{d}_1$ and $\hat{d}_2$. We will begin by showing that $\hat{d}_1$ is more efficient than $\hat{d}_2$ for the M/M/2 queue. Later in this section we will consider more general systems by means of simulation.

Consider an M/M/2 queue with Poisson arrival rate λ and identical servers with exponential service rate μ . A normal M/M/2 busy cycle begins when a customer arrives to find the system completely empty. However, in order to facilitate computation we are going to consider a modified busy cycle. The M/M/2 queue will also regenerate itself at those epochs when a customer arrives to find exactly two customers in the system. Let B'_c be the time between two such

epochs. Similarly, let N'_c be the number of customers served in B'_c and let D'_c be the total delay of these N'_c customers. Consider the following estimators of d :

$$(1.3) \quad \hat{d}'_1 = \frac{\bar{D}'_c(k)}{\bar{N}'_c(k)}$$

and

$$(1.4) \quad \hat{d}'_2 = \frac{1}{\lambda} \frac{\bar{D}'_c(k)}{\bar{B}'_c(k)},$$

where k is the length of the simulation in modified busy cycles.

We will see later in this section that the first and second moments of D'_c , N'_c , and B'_c are finite. Furthermore, by following the argument of Theorem 14.4 of Chung (1967), it can be shown that the finiteness of the above moments implies the finiteness of the first and second moments of the unprimed r.v.'s D_c , N_c , and B_c . It follows from Chapter 1 that

$$(1.5) \quad d = \frac{E(D'_c)}{E(N'_c)} = \frac{1}{\lambda} \frac{E(D'_c)}{E(B'_c)}.$$

We would like to show that comparing the VAD's of $\hat{d}'_1$ and $\hat{d}'_2$ is equivalent to comparing the VAD's of $\hat{d}_1$ and $\hat{d}_2$. Let V_1 be as defined by Equation (3.6) of Chapter 1. Define V'_1 , V_2 , and V'_2 analogously. It follows from Iglehart (1973) that

$$(1.6) \quad E(N'_c)V'_1 = E(N_c)V_1$$

and

$$(1.7) \quad E(B'_c)V'_2 = E(B_c)V_2.$$

Furthermore, $E(N_c) = \lambda E(B_c)$ and $E(N'_c) = \lambda E(B'_c)$. It follows that

$$(1.8) \quad \frac{VAD(d'_1)}{VAD(d'_2)} = \frac{VAD(d_1)}{VAD(d_2)}.$$

Therefore, for the remainder of our study of the M/M/2 queue, we will concentrate our attention on d'_1 and d'_2 .

In order to evaluate the VAD's of d'_1 and d'_2 , we must be able to compute the first and second moments of D'_c , N'_c , and B'_c . Since the techniques involved are similar to before, we will not give the complete details. Consider an arbitrary modified busy cycle, B'_c , as shown in Figure 4.1. The shaded area represents D'_c . This area is also the total delay during the period of time when the system goes from 3 to 2 customers present. However, the behavior of the system during this period of time is, for our purposes, the same as the behavior of an M/M/1 queue, with arrival and service rates λ and 2μ , during the period of time it goes from 2 customers to 1 customer present. Let D''_c , N''_c , and B'' denote the total delay per cycle, number of customers per cycle, and length of a busy period for the aforementioned M/M/1 queue. Furthermore, let T_{23} denote the time required for the system to go from 2 back to 3 customers present, and let N_{23} denote the number of customers who complete service during T_{23} . Note that the system may visit states 1 and 0 before returning to 3. Therefore,

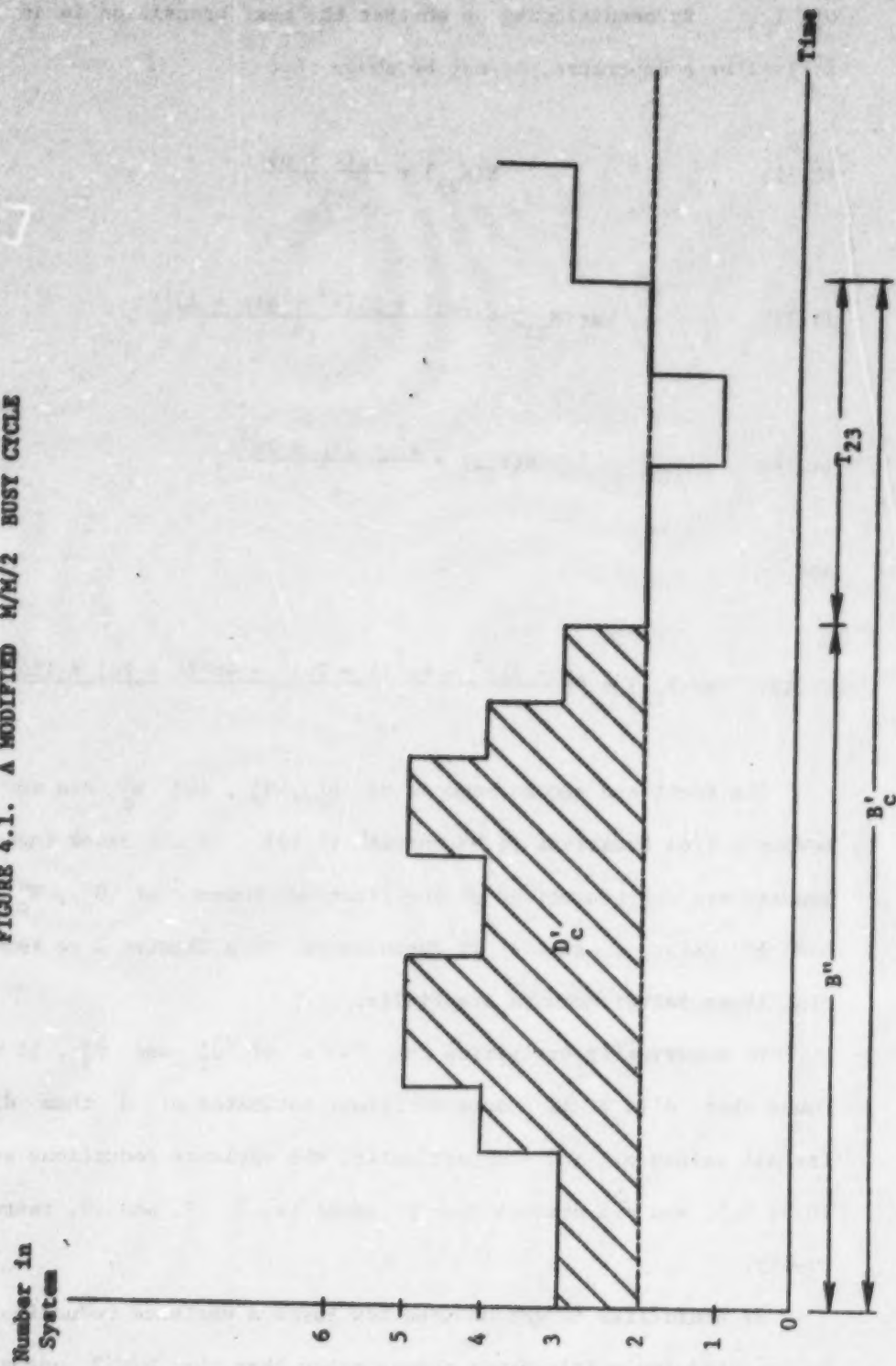
$$(1.9) \quad D'_c = D''_c + B'',$$

$$(1.10) \quad N'_c = N''_c + N_{23},$$

and

$$(1.11) \quad B'_c = B'' + T_{23}.$$

FIGURE 4.1. A MODIFIED M/M/2 BUSY CYCLE



In addition, N_c'' is independent of N_{23} and B'' is independent of T_{23} . By conditioning on whether the next transition is an arrival or a departure, it may be shown that

$$(1.12) \quad E(N_{23}) = \frac{2\mu(\lambda + \mu)}{\lambda^2},$$

$$(1.13) \quad \text{Var}(N_{23}) = \frac{2\mu(\lambda + \mu)(\lambda^2 + 4\lambda\mu + 2\mu^2)}{\lambda^4},$$

$$(1.14) \quad E(T_{23}) = \frac{\lambda^2 + 2\lambda\mu + 2\mu^2}{\lambda^3},$$

and

$$(1.15) \quad \text{Var}(T_{23}) = \frac{(\lambda + 2\mu)^4 - 4\mu^2(\lambda + 2\mu)^2 - 4\mu^3(\lambda + 2\mu) + 12\mu^4}{\lambda^6}.$$

The first and second moments of D_c' , N_c' , and B_c' can now be computed from Equations (1.9) through (1.15). In all cases these moments are only functions of the first two moments of D_c'' , N_c'' , and B'' (also λ and μ). Furthermore, from Chapter 2 we know that these latter moments are finite.

By numerically evaluating the VAD's of $\hat{d}_1'$ and $\hat{d}_2'$, it was found that $\hat{d}_1'(\hat{d}_1)$ is a more efficient estimator of d than $\hat{d}_2'(\hat{d}_2)$ for all values of ρ . In particular, the variance reductions were 10.9, 9.5, and 4.3 percent for ρ equal to .5, .7, and .9, respectively.

We would like to get an idea how large a variance reduction can be expected for multi-server queues other than the M/M/2 queue. In Chapters 2 and 3 we found that the variance reduction resulting from

using $\hat{d}_1$ rather than $\hat{d}_2$, increased as the interarrival distribution becomes more irregular, and increased as the service distribution becomes more regular. With this in mind we simulated a GI/G/2 queue with $\rho = .5$, hyperexponential interarrival distribution (the same one used in Chapter 3), and 4-Erlang service distribution. We made 100 independent runs each of length 100 normal busy cycles. We found the variance reduction from using $\hat{d}_1$ rather than $\hat{d}_2$ to be 46.9 percent. This is much larger than what we found for the M/M/2 queue, but consistent with our results for single server queues with the same distributions.

We also found that the average number of customers per cycle was 18.9. Thus, the above comparison was based on almost 200,000 customers.

4.2 Estimation of w for GI/G/s queues ($s \geq 1$)

In this section we consider the efficient estimation of mean waiting time, w , for arbitrary GI/G/s queues ($s \geq 1$). Suppose we run a simulation until n customers have been served. The usual classical estimator of d is given by

$$(5.1) \quad \tilde{w}_1 = \frac{\sum_{i=1}^n (D_i + S_i)}{n} = \bar{d}_1 + \bar{S},$$

where $\bar{S}$ is the sample mean of the i.i.d. sequence $\{S_i\}$. However, w can also be estimated by (see Section 1.4)

$$(5.2) \quad \tilde{w}_2 = \bar{d}_1 + E(S).$$

Our objective is to show that the variance of $\tilde{w}_2$ is less than the variance of $\tilde{w}_1$ for all values of n .

We begin with a characterization of GI/G/s queues which is

due to Kiefer and Wolfowitz (1955,1956). For a fixed realization of the queueing system, think of assigning customers when they arrive to the server who will eventually serve them. A customer is always assigned to the first available server which has the smallest index.

Let $\underline{D}_n = (D_{n1}, D_{n2}, \dots, D_{ns})$ be the vector-valued process whose components denote the workload of the server with the i^{th} lightest load prior to the arrival of the n^{th} customer. Thus D_{n1} is the actual delay in queue of the n^{th} customer. The sequence $\{\underline{D}_n : n \geq 1\}$ can be generated recursively as follows:

$$(5.3) \quad \underline{D}_1 = \underline{0}$$

and

$$(5.4) \quad \underline{D}_{n+1} = [F(\underline{D}_n + \underline{S}_n - \underline{A}_n)]^+,$$

where $\underline{S}_n = (S_n, 0, \dots, 0)$, $\underline{A}_n = (A_n, A_n, \dots, A_n)$, $F: \mathbb{R}^s \rightarrow \mathbb{R}^s$ is a function which arranges the components of its argument in increasing order, and $[\underline{X}]^+ = (x_1^+, x_2^+, \dots, x_s^+)$ for $\underline{X} = (x_1, x_2, \dots, x_s) \in \mathbb{R}^s$.

We now state a lemma whose proof can be found in Esary, Proschan, and Walkup (1967).

Lemma 2.1:

Let $X_1, X_2, \dots, X_n$ be independent r.v.'s. If the functions f and g are nondecreasing in each argument, then $\text{Cov}[f(X_1, X_2, \dots, X_n), g(X_1, X_2, \dots, X_n)] \geq 0$.

We now come to the main result of this section.

Theorem 2.2:

For all sample sizes n , we have

$$(5.5) \quad \text{Var}(\tilde{w}_1) \geq \text{Var}(\tilde{w}_2)$$

and

$$(5.6) \quad \text{MSE}(\bar{W}_1) \geq \text{MSE}(\bar{W}_2) ,$$

equality holding in (5.5) and (5.6) if and only if $\text{Var}(\bar{S}) = 0$.

Proof:

It follows from (5.1) and (5.2) that

$$(5.7) \quad \text{Var}(\bar{W}_2) = \text{Var}(\bar{d}_1)$$

and

$$(5.8) \quad \text{Var}(\bar{W}_1) = \text{Var}(\bar{d}_1) + 2\text{Cov}(\bar{d}_1, \bar{S}) + \text{Var}(\bar{S}) .$$

Furthermore,

$$(5.9) \quad \text{Cov}(\bar{d}_1, \bar{S}) = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n \text{Cov}(D_i, S_j)$$

and

$$(5.10) \quad \text{Var}(\bar{S}) = \frac{\text{Var}(\bar{S})}{n} \geq 0 .$$

It will be sufficient to show that $\text{Cov}(\bar{d}_1, \bar{S}) \geq 0$. We will do this by showing that $\text{Cov}(D_i, S_j) \geq 0$ for every i and j .

For a GI/G/s queue the delay of a customer is independent of his and all future service times. Therefore, $\text{Cov}(D_i, S_j) = 0$ for $i = 1, 2, \dots, n$ and $j = i, i+1, \dots, n$. Furthermore, it is clear that D_i and S_j ($j = 1, 2, \dots, i-1$) can be considered nondecreasing functions of the independent r.v.'s $S_1, S_2, \dots, S_{i-1}, -A_1, -A_2, \dots, -A_{i-1}$; that is, there exist nondecreasing functions f and g_j ($j = 1, 2, \dots, i-1$) such that

$$(5.11) \quad D_1 = f(s_1, s_2, \dots, s_{i-1}, -A_1, -A_2, \dots, -A_{i-1})$$

and

$$(5.12) \quad s_j = g_j(s_1, s_2, \dots, s_{i-1}, -A_1, -A_2, \dots, -A_{i-1})$$

for $j = 1, 2, \dots, i-1$. It follows from Lemma 2.1 that

$\text{Cov}(D_1, s_j) \geq 0$ for $i = 2, 3, \dots, n$ and $j = 1, 2, \dots, i-1$.

Thus, (5.5) is proven. Furthermore, it is clear from (5.9) and (5.10) that equality will hold in (5.5) if and only if $\text{Var}(S) = 0$.

Let us now consider (5.6). It is easy to see that

$$(5.13) \quad \text{Bias}^2(\tilde{w}_1) = [E(\tilde{w}_1) - w]^2 = [E(\tilde{d}_1) + E(\bar{s}) - d - E(s)]^2 = \\ [E(\tilde{d}_1) - d]^2$$

and

$$(5.14) \quad \text{Bias}^2(\tilde{w}_2) = [E(\tilde{w}_2) - w]^2 = [E(\tilde{d}_1) + E(s) - d - E(s)]^2 = \\ [E(\tilde{d}_1) - d]^2.$$

The second part of the theorem now follows.

CHAPTER 5

COMPARISON OF THE REGENERATIVE AND CLASSICAL METHODS FOR
SMALL SAMPLE SIZES5.1 Comparison of $\hat{d}_1$ and $\tilde{d}_1$

We have shown that a regenerative estimator and its comparable classical estimator have the same asymptotic normal distributions. Therefore, for very large samples it makes little difference which method is used. However, in many real world situations we may not have the luxury of such large sample sizes. Therefore, it is important to compare the two methods for small samples.

Suppose we simulate a queueing system for m busy cycles. The usual regenerative estimator of d is given by

$$(1.1) \quad \hat{d}_1 = \frac{\bar{D}_c(m)}{\bar{N}_c(m)}.$$

Suppose we also simulate the same queueing system until n customers have been served, where n is chosen such that $n = mE(N_c)$. The usual classical estimator of d is given by

$$(1.2) \quad \tilde{d}_1 = \frac{\sum_{i=1}^n D_i}{n}.$$

Thus, on the average, the two estimators are based on the same number of customers. Our objective is to empirically compare $\hat{d}_1$ and $\tilde{d}_1$ for small sample sizes on the basis of their MSE's.

Denote the r.v. $(\hat{d}_1 - d)^2$ by X_1 , and the r.v. $(\tilde{d}_1 - d)^2$ by X_2 . Thus, we will compare $E(X_1)$ and $E(X_2)$.

Suppose that we make k independent simulation runs of our queueing system. It is possible, and also efficient, to get both

a regenerative and classical estimate from a single run. A run is continued until either n busy cycles are completed, or until n customers have been served. At this point the appropriate estimate is made. The simulation is then continued until the other requirement is met, at which point the second estimate is made. Let X_{ij} be j^{th} ($j = 1, 2, \dots, k$) observation of X_i . An unbiased estimate of $E(X_i)$ is given by

$$(1.3) \quad \bar{X}_i = \frac{\sum_{j=1}^k X_{ij}}{k} \quad (i = 1, 2) .$$

In order to compare $\hat{d}_1$ and $\bar{X}_1$ as stated, we need to consider a queueing system for which d and $E(N_c)$ are known. We have a number of choices, but have chosen the M/M/1 queue because it is easy and inexpensive to simulate. In particular, we considered the following two M/M/1 queues:

$$(1.4) \quad \rho = 0.5, E(A) = 1, E(S) = 0.5, d = 0.5$$

and

$$(1.5) \quad \rho = 0.9, E(A) = 1, E(S) = 0.9, d = 8.1 .$$

For these systems $E(N_c)$ is equal to 2 and 10, respectively. In each case we made 1000 runs each of length 100 cycles. Thus, n was equal to 200 and 1000 customers, respectively. The results of these simulation runs are given in Table 5.1. In both cases the regenerative estimator appears to have smaller MSE than the classical estimator.

We would like to determine whether the observed differences are real or whether they were caused by sampling fluctuations. It will be sufficient to test the hypothesis

TABLE 5.1. COMPARISON OF $MSE(\hat{d}_1)$ AND $MSE(\hat{d}'_1)$

ρ	$\bar{x}_1$	$\bar{x}_2$	s_{11}	s_{22}	s_{12}	$\frac{s_{12}}{\sqrt{s_{11}s_{22}}}$	t
0.5	.03439	.03736	.00704	.01426	.00944	.94	1.90
0.9	14.552	17.523	805.15	3431.23	1215.89	.73	2.21

$$(1.6) \quad H_0 : E(X_1) = E(X_2)$$

against the alternative

$$(1.7) \quad H_1 : E(X_1) \neq E(X_2) .$$

Let

$$(1.8) \quad \sigma^2 = \text{Var}(X_1) - 2\text{Cov}(X_1, X_2) + \text{Var}(X_2) .$$

If H_0 is true, then $\bar{X}_2 - \bar{X}_1$ has mean 0, variance $\frac{\sigma^2}{k}$, and

$$(1.9) \quad \frac{k^{1/2}(\bar{X}_2(k) - \bar{X}_1(k))}{\sigma} \xrightarrow[k \rightarrow \infty]{D} N(0,1) .$$

Since σ^2 is unknown it can be estimated by

$$(1.10) \quad S^2(k) = S_{11}(k) - 2S_{12}(k) + S_{22}(k) ,$$

where

$$(1.11) \quad S_{ij}(k) = \frac{\sum_{l=1}^k (X_{il} - \bar{X}_i)(X_{jl} - \bar{X}_j)}{k-1} \quad (i = 1, 2; j = 1, 2) .$$

Furthermore, it is easy to show that

$$(1.12) \quad \lim_{k \rightarrow \infty} S(k) = \sigma \quad \text{w.p. 1} .$$

Therefore, it follows from Lemma 3.1 of Chapter 1 that

$$(1.13) \quad \frac{k^{1/2}(\bar{X}_2(k) - \bar{X}_1(k))}{S(k)} \xrightarrow[k \rightarrow \infty]{D} N(0,1) .$$

Hence, for large k

$$(1.14) \quad t = \frac{k^{1/2}(\bar{x}_2(k) - \bar{x}_1(k))}{s(k)}$$

will be approximately normally distributed with mean 0 and variance 1.

Considering the hypothesis H_1 , the appropriate critical region is of the form reject H_0 if $|t| > t_0$. The values of t for the two simulation experiments were 1.90 and 2.21 for $\rho = 0.5$ and 0.9, respectively. Thus, the observed levels of significance for the two experiments were 0.057 and 0.027, where by observed level of significance we mean the smallest significance level for which a test of the above form would reject H_0 on the basis of the observed data. It appears that the regenerative estimator is superior in both cases. However, it is believed that this is an area which warrants further research.

APPENDIX

A.1 Moments for the M/G/1 Queue

We now consider the computation of the moments required to evaluate the VAD's of the estimators considered in Chapter 2. In order to understand the basic technique that is involved, we begin with a detailed discussion of how to compute the expected length of a busy period, $E(B)$, for the M/G/1 queue. We will condition on the following r.v.'s:

S = service time of the customer who begins the busy period,

N = number of customers who arrive during S .

The length of a busy period conditioned on N and S is given by

$$(1.1) \quad (B \mid N, S) = S + \sum_{i=1}^N B_i,$$

where B_i ($i = 1, 2, \dots, N$) is the length of the busy period started by the i^{th} customer who arrived during S . A word of explanation is in order at this point. We are conditioning on N and S so that the busy period begins with a customer who has a service time of S .

Furthermore, during this service time N additional customers arrive.

At the epoch the original customer finishes his service, the first of the N remaining customers enters service. Let us set the other

$N - 1$ customers aside for the moment. Because of the Poisson arrival stream, the entering customer finds the system in exactly the same state as the original customer did when he entered the system. Thus, the time required to service the customer who is now entering service, the customers who arrive during his service, the customers who arrive

during these services, etc., call it B_1 , has exactly the same distribution as the original busy period, B . After B_1 has been completed there are $N - 1$ customers remaining in the system. The first of these enters service and starts a busy period, call it B_2 , with the same distribution as B . Furthermore, B_2 is independent of B_1 . Continuing in this manner, it is clear that we have a total of N i.i.d. busy periods with the same distribution as B . Therefore, taking the expected value of both sides of (1.1) we get

$$(1.2) \quad E(B \mid N, S) = S + NE(B).$$

Since N given S is a Poisson r.v. with mean λS , if we uncondition with respect to N in (1.2) we get

$$(1.3) \quad E(B \mid S) = S + \lambda SE(B).$$

Finally, unconditioning with respect to S in (1.3) we get

$$(1.4) \quad E(B) = E(S) + \lambda E(S)E(B)$$

or

$$(1.15) \quad E(B) = \frac{E(S)}{1 - \rho}.$$

Since

$$(1.16) \quad (B^2 \mid N, S) = \left(S + \sum_{i=1}^N B_i \right)^2 = S^2 + 2S \sum_{i=1}^N B_i + \sum_{i=1}^N \sum_{j=1}^N B_i B_j,$$

it is easy to show that

$$(1.17) \quad E(B^2) = \frac{E(S^2)}{(1 - \rho)^3}.$$

The moments of N_c , B_c , and $B + \frac{1}{\lambda}$ can be computed in a similar manner to those of B . They are as follows:

$$(1.18) \quad E(N_c) = \frac{1}{1 - \rho},$$

$$(1.19) \quad E(N_c^2) = E^2(N_c)[1 + \rho + \lambda^2 E(N_c)E(S^2)],$$

$$(1.20) \quad E(B_c) = \frac{E(N_c)}{\lambda},$$

$$(1.21) \quad E(B_c^2) = E^3(N_c)E(S^2) + \frac{2E(N_c)}{\lambda^2},$$

$$(1.22) \quad E\left(B + \frac{1}{\lambda}\right) = \frac{E(N_c)}{\lambda},$$

and

$$(1.23) \quad E\left\{\left(B + \frac{1}{\lambda}\right)^2\right\} = E^3(N_c)E(S^2) + \frac{[2E(N_c) - 1]}{\lambda^2}.$$

We now come to the moments of D_c . The total delay per cycle conditioned on N and S is given by

$$(1.24) \quad (D_c | N, S) = \sum_{i=1}^N T_i + \sum_{i=1}^N (N - i)B_i + \sum_{i=1}^N D_{c_i},$$

where T_i ($i = 1, 2, \dots, N$) is the amount of time the i^{th} arrival during S has to wait until the service time of the original customer, S , has been completed. The second term on the r.h.s. of (1.24) results from the delay of the $N - 1$ customers who were set aside during B_1 , the $N - 2$ customers who were set aside during B_2 , etc. In the last term D_{c_i} represents the total delay of all

customers who were served during the busy period initiated by the i^{th} arrival during S . Furthermore, the $\{D_{c_i}\}$ are independent and all have the same distribution as D_c . Using (1.24) it can be shown that

$$(1.25) \quad E(D_c) = \frac{\lambda E^2(N_c) E(S^2)}{2},$$

$$(1.26) \quad E(D_c^2) = E^3(N_c) \left[\frac{\lambda^2 E(S^4)}{4} + \frac{\lambda(1+\rho) E(S^3)}{3} + \frac{4\lambda^3 E(N_c) E(S^3) E(S^2)}{3} + \frac{5\lambda^4 E^2(N_c) E^3(S^2)}{4} + \lambda^2 E(N_c) \left(1 + \frac{\rho}{2}\right) E^2(S^2) \right],$$

$$(1.27) \quad E(D_c B) = E^3(N_c) \left[\frac{\lambda E(S^3)}{2} + \lambda^2 E(N_c) E^2(S^2) + \frac{\rho E(S^2)}{2} \right],$$

and

$$(1.28) \quad E(D_c N_c) = E^3(N_c) \left[\frac{\lambda^2 E(S^3)}{2} + \lambda^3 E(N_c) E^2(S^2) + \lambda \left(1 + \frac{\rho}{2}\right) E(S^2) \right].$$

The moments given by (1.26) and (1.27) may be found in Daley and Jacobs (1969).

We would now like to compute moments for V_c , analogous to the ones that we computed for D_c . We will condition on the service time of the first customer, S_0 ; the number of customers, N , who arrive during S_0 ; and the service times of the N customers; $S_1, S_2, \dots, S_N$. We have the following relationship:

$$\begin{aligned}
 (1.29) \quad (V_c \mid s_1, s_2, \dots, s_N, N, s_0) &= \frac{s_0^2}{2} + \sum_{i=1}^N T_i s_i \\
 &+ \sum_{i=1}^{N-1} \left(\sum_{j=i+1}^N s_j \right) (B_i \mid s_i) + \sum_{i=1}^N (V_{c_i} \mid s_i).
 \end{aligned}$$

The first term on the r.h.s. of (1.29) is the contribution of the original customer while he was in service. The $T_i s_i$ in the second term is the contribution of the i^{th} customer (who arrived during s_0) while waiting for s_0 to be completed. The $(B_i \mid s_i)$ is the length of the busy period initiated by i^{th} arrival, conditioned on the service time of the i^{th} arrival. Thus, $\left(\sum_{j=i+1}^N s_j \right) (B_i \mid s_i)$ is the contribution of the remaining $N - i$ customers while waiting for B_i to be completed. The $(V_{c_i} \mid s_i)$ in the fourth term is the contribution of all the customers served during B_i , conditioned on the service time of the customer (the i^{th} arrival) who begins B_i .

After lengthy calculations, it can be shown that

$$(1.30) \quad E(V_c) = \frac{E^2(N_c)E(S^2)}{2},$$

$$(1.31) \quad E(V_c^2) = E^3(N_c) \left[\frac{E(S^4)}{4} + \frac{4\lambda E(N_c)E(S^3)E(S^2)}{3} + \frac{5\lambda^2 E^2(N_c)E^3(S^2)}{4} \right],$$

$$(1.32) \quad E(V_c B) = E^3(N_c) \left[\frac{E(S^3)}{2} + \lambda E(N_c)E^2(S^2) \right],$$

and

$$(1.33) \quad E(V_c N_c) = E^3(N_c) \left[\frac{\lambda E(S^3)}{2} + \lambda^2 E(N_c)E^2(S^2) + \frac{(1 + \rho)E(S^2)}{2} \right].$$

The moments of V_c^q can be computed in a similar manner to those of V_c . They are as follows:

$$(1.34) \quad E(V_c^q) = \frac{\rho E^2(N_c) E(S^2)}{2},$$

$$(1.35) \quad E\left\{(V_c^q)^2\right\} = E^3(N_c) \left\{ \frac{\rho^2 E(S^4)}{4} + \frac{\lambda E(N_c)}{3} \left[1 + \frac{3\rho(1+\rho)}{2} \right] E(S^3) E(S^2) + \frac{\lambda^2 E^2(N_c) [1 + (1+\rho)^2] E^3(S^2)}{4} \right\},$$

$$(1.36) \quad E(V_c^{qb}) = E^3(N_c) \left[\frac{\rho E(S^3)}{2} + \frac{\lambda(1+\rho) E(N_c) E^2(S^2)}{2} \right],$$

and

$$(1.37) \quad E(V_c^{qN_c}) = E^3(N_c) \left[\frac{\lambda \rho E(S^3)}{2} + \frac{\lambda^2(1+\rho) E(N_c) E^2(S^2)}{2} + \rho E(S^2) \right].$$

A.2 Moments for Single Server Queues with Batch Poisson Arrivals

In this section we consider the computation of the moments required to evaluate the VAD's of the estimators considered in Section 3.2. Namely, the estimators of d for single server queues with Poisson arrivals in batches of size k . Of course, for $k = 1$ the results of this section are identical with those in Section A.1.

A busy period will now begin when a batch of k customers arrives to find the system empty. In order to compute the required moments, we condition on the service times of the k customers who begin a busy period; $S_1, S_2, \dots, S_k$; and the number of batches of customers, N , which arrive during the time required to service these k customers. Since the calculations are similar, but more

laborious than those of the last section, we will only give the final expressions for the desired moments. Let $Y = \sum_{i=1}^k S_i$. The moments follow:

$$(2.1) \quad E(N_c) = \frac{k}{1 - \rho},$$

$$(2.2) \quad E(N_c^2) = E^2(N_c) \left[1 + \rho + \frac{\lambda^2 E(N_c) E(Y^2)}{k} \right],$$

$$(2.3) \quad E(B) = \frac{E(N_c) E(Y)}{k},$$

$$(2.4) \quad E(B^2) = \frac{E^3(N_c) E(Y^2)}{k^3},$$

$$(2.5) \quad E(B_c) = E(B) + \frac{1}{\lambda},$$

$$(2.6) \quad E(B_c^2) = E(B^2) + \frac{2E(B_c)}{\lambda},$$

$$(2.7) \quad E(D_c) = \frac{E(N_c) [(k-1)E(Y) + \lambda E(N_c) E(Y^2)]}{2k}.$$

Let

$$(2.8) \quad a = \frac{\lambda^2 E^2(N_c)}{4},$$

$$(2.9) \quad b = \frac{4\lambda^3 E^3(N_c) E(Y^2)}{3k} + \frac{\lambda^2 E^2(N_c) E(Y)}{3} + \frac{\lambda E^2(N_c)}{3} \\ + \frac{(k-1)\lambda^2 E^2(N_c) E(Y)}{2k} + \frac{(k-1)\lambda E(N_c)}{2},$$

$$(2.10) \quad c = \frac{5\lambda^4 E^4(N_c) E^2(Y^2)}{4k^2} + \left[\frac{(3k-2)\lambda^3 E^3(N_c) E(Y)}{2k^2} + \frac{\lambda^2 E^3(N_c)}{k} + \frac{(k-1)\lambda^2 E^2(N_c)}{k} \right] E(Y^2),$$

$$(2.11) \quad d = \left[\frac{(k-1)^2 \lambda^2 E^2(N_c) E(Y)}{4k^2} + \frac{(k-1)^2 \lambda E(N_c)}{2k} + \frac{(k-1)\lambda E^2(N_c)}{2k} \right] E(Y),$$

and

$$(2.12) \quad e = \frac{k(k-1)(2k-1)E(S^2)}{6} + \frac{k(k-1)(k-2)(3k-1)E^2(S)}{12}.$$

Then

$$(2.13) \quad E(D_c^2) = \frac{E(N_c)}{k} [aE(Y^4) + bE(Y^3) + (c+d)E(Y^2) + e].$$

Let

$$(2.14) \quad f = \frac{\lambda^2 E(N_c)}{2},$$

$$(2.15) \quad g = \frac{\lambda^3 E^2(N_c) E(Y^2)}{k} + \frac{(2k-1)\lambda^2 E(N_c) E(Y)}{2k} + \lambda E(N_c) + \frac{(k-1)\lambda}{2},$$

and

$$(2.16) \quad h = \frac{(k-1)E(Y)}{2}.$$

Then

$$(2.17) \quad E(D_c N_c) = \frac{E^2(N_c)}{k} [fE(Y^3) + gE(Y^2) + h] .$$

Let

$$(2.18) \quad i = \frac{\lambda E(N_c)}{2}$$

and

$$(2.19) \quad j = \frac{\lambda^2 E^2(N_c) E(Y^2)}{k} + \frac{(2k - 1) \lambda E(N_c) E(Y)}{2k} + \frac{k - 1}{2} .$$

Then

$$(2.20) \quad E(D_c B) = \frac{E^2(N_c)}{k^2} [iE(Y^3) + jE(Y^2)] .$$

In order to numerically evaluate the above moments, it is necessary to know the first four moments of Y . These moments follow:

$$(2.21) \quad E(Y) = kE(S) ,$$

$$(2.22) \quad E(Y^2) = kE(S^2) + k(k - 1)E^2(S) ,$$

$$(2.23) \quad E(Y^3) = kE(S^3) + 3k(k - 1)E(S^2)E(S) + k(k - 1)(k - 2)E^3(S) ,$$

and

$$(2.24) \quad \begin{aligned} E(Y^4) = & kE(S^4) + 4k(k - 1)E(S^3)E(S) + 3k(k - 1)E^2(S^2) \\ & + 6k(k - 1)(k - 2)E(S^2)E^2(S) + k(k - 1)(k - 2)(k - 3)E^4(S) . \end{aligned}$$

REFERENCES

- [1] Blomqvist, N., "On the Transient Behavior of the GI/G/1 Waiting-Times," Skand. Akt. Tidskr., pp. 118-129, (1970).
- [2] Brumelle, S. L., "Some Inequalities for Parallel-Server Queues," Operations Research, Vol. 19, pp. 402-413, (1971).
- [3] Chung, K. L., MARKOV CHAINS WITH STATIONARY TRANSITION PROBABILITIES, Springer-Verlag, Berlin, (1967).
- [4] Chung, K. L., A COURSE IN PROBABILITY THEORY, Harcourt, Brace, and World, New York, (1968).
- [5] Conway, R. W., "Some Tactical Problems in Digital Simulation," Management Science, Vol. 10, pp. 47-61, (1963).
- [6] Cox, D. R. and W. L. Smith, QUEUES, Methuen, London, (1961).
- [7] Crane, M. A. and D. L. Iglehart, "A New Approach to Simulating Stable Stochastic Systems, I: General Multi-Server Queues," Technical Report 86-1, Control Analysis Corporation, Palo Alto, California, (1972a).
- [8] Crane, M. A. and D. L. Iglehart, "Confidence Intervals for the Ratio of Two Means with Application to Simulations," Technical Report 86-2, Control Analysis Corporation, Palo Alto, California, (1972b).
- [9] Daley, D. J. and D. R. Jacobs, Jr., "The Total Waiting Time in a Busy Period of a Stable Single-Server Queue, II," J. Appl. Prob., Vol. 6, pp. 565-572, (1969).
- [10] Esary, J. D., F. Proschan and D. W. Walkup, "Association of Random Variables, with Applications," Ann. Math. Stat., Vol. 38, pp. 1466-1474, (1967).
- [11] Feller, W., AN INTRODUCTION TO PROBABILITY THEORY AND ITS APPLICATIONS, Vol. 2, Wiley, New York, (1966).
- [12] Fishman, G. S. and P. J. Kiviat, "The Analysis of Simulation-Generated Time Series," Management Science, Vol. 13, pp. 525-557, (1967).
- [13] Fishman, G. S., "Output Analysis for Queueing Simulations," Technical Report 56, Yale University, (1972).
- [14] Iglehart, D. L., "Functional Limit Theorems for the Queue GI/G/1 in Light Traffic," Adv. Appl. Prob., Vol. 3, pp. 269-281, (1971).
- [15] Iglehart, D. L., "Discrete-Event Simulations with Two or More Sequences of Regeneration Times," Technical Report 86-9, Control Analysis Corporation, Palo Alto, California, (1973).

- [16] Kiefer, J. and J. Wolfowitz, "On the Theory of Queues with Many Servers," Trans. Amer. Math. Soc., Vol. 78, pp. 1-18, (1955).
- [17] Kiefer, J. and J. Wolfowitz, "On the Characteristics of the General Queueing Process, with Applications to Random Walk," Ann. Math. Stat., Vol. 27, pp. 147-161, (1956).
- [18] Little, J.D.C., "A Proof of the Queueing Formula: $L = \lambda w$," Operations Research, Vol. 9, pp. 383-388, (1961).
- [19] Lindley, D. V., INTRODUCTION TO PROBABILITY AND STATISTICS, Part 1, Cambridge University Press, (1969).
- [20] Ross, S. M., APPLIED PROBABILITY MODELS WITH OPTIMIZATION APPLICATIONS, Holden-Day, San Francisco, (1970).
- [21] Sclove, S. L., G. Simons and J. Van Ryzin, "Further Remarks on the Expectation of the Reciprocal of a Positive Random Variable," The Amer. Stat., Vol. 21, pp. 33-34, (1967).
- [22] Smith, W. L., "Regenerative Stochastic Processes," Proc. Royal Soc., Series A, Vol. 232, pp. 6-31, (1955).
- [23] Whitt, W., "Embedded Renewal Processes in the GI/G/s Queue," J. Appl. Prob., Vol. 9, pp. 650-658, (1972).
- [24] Wolff, R. W., "Work-Conserving Priorities," J. Appl. Prob., Vol. 7, pp. 327-337, (1970).

GLOSSARY

- A_i the time between the arrivals of the i^{th} and $i + 1^{\text{st}}$ customers
 $E(A)$ the mean interarrival time
 λ $\frac{1}{E(A)}$, the arrival rate of customers
 S_i the service time of the i^{th} customer
 $E(S^r)$ the r^{th} moment of the service distribution
 μ $\frac{1}{E(S)}$, the service rate
 ρ the traffic intensity
 D_i the delay in queue of the i^{th} customer
 $N_q(t)$ the number of customers in the queue at epoch t
 $V^q(t)$ the work in queue at epoch t , i.e., the sum of the service times of the customers in queue
 $V^s(t)$ the work in service at epoch t , i.e., the sum of the remaining service times of the customer(s) in service
 $V(t)$ $V^q(t) + V^s(t)$, the work in system at epoch t
 d the mean delay in queue
 Q the time average number of customers in queue
 w $d + E(S)$, the mean waiting time in system
 $E(V^q)$ the time average amount of work in queue
 $E(V^s)$ the time average amount of work in service
 $E(V)$ $E(V^q) + E(V^s)$, the time average amount of work in system
 B the length of a busy period
 I the length of an idle period
 B_c the length of a busy cycle, a busy period plus the succeeding idle period
 N_c the number of customers served in a busy cycle (period)
 D_c the total delay of all customers served in a busy cycle

$V_c^q \int_0^{B_c} V^q(t) dt$, the total work in queue per cycle

$V_c \int_0^{B_c} V(t) dt$, the total work per cycle